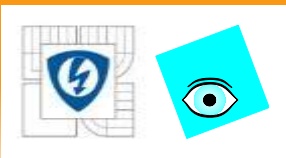


Statistická analýza jednorozměrných dat

Prof. RNDr. Milan Meloun, DrSc.
Univerzita Pardubice, Pardubice

31.ledna 2011

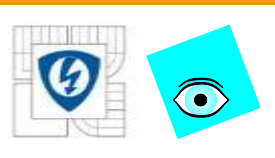
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky.



Kapitola 2.1

EDA – Exploratory Data Analysis (Tuckey, Chambers)

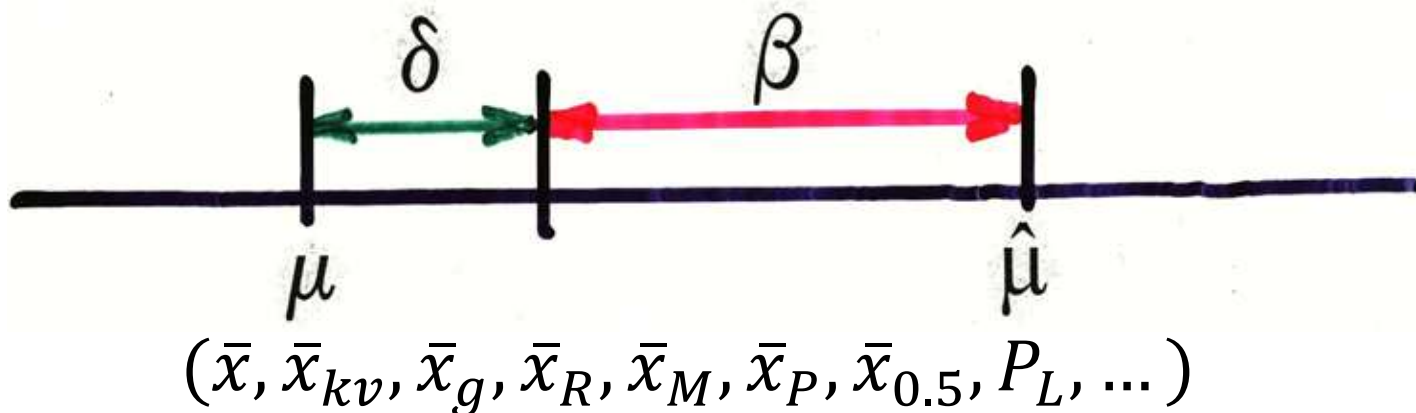
PRŮZKUMOVÁ ANALÝZA DAT

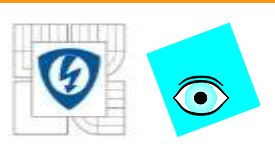


Systematické chyby

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

δ **odchylka** špatným experimentem, nezvládnutou
strategií





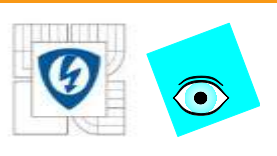
Náhodné chyby

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



Šíře i. s. závisí na:

1. Četnosti n
2. Velikosti S (směrodatná odchylka, nejistota, zákon propagace).
3. Statistické jistotě 90%, 95%, 99%, ...

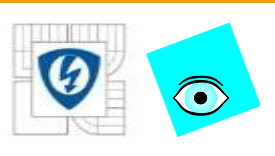


1. Průzkumová analýza dat:

Diagnostické grafy: stupeň symetrie rozdělení
 lokální koncentrace dat
 vybočující data

2. Ověření předpokladů výběru dat:

Diagnostické testy: ověření normality
 ověření nezávislosti
 ověření homogenity
 určení minimální četnosti

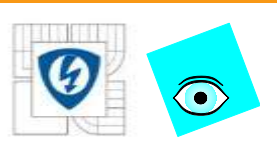


3. Transformace dat:

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

2. Parametry polohy, rozptýlení a tvaru:

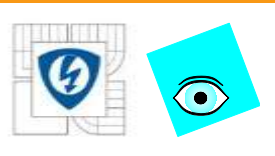
Analýza 1 výběru: klasické odhady (průměr, rozptyl)
 robustní odhady (medián, ořezané průměry,
 winsorizovaný rozptyl, interkvantilové rozpětí)
 adaptivní odhady



Průzkumová analýza dat (EDA)

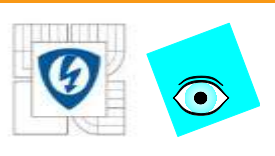
Cílem průzkumové analýzy dat je **nalezení zvláštností statistického chování dat** a **ověření jejich předpokladů** pro následné zpracování „klasickými“ statistickými metodami.

EDA – **E**xploratory **D**ata **A**nalysis (Tuckey, Chambers)



Průzkumová analýza dat (EDA)

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



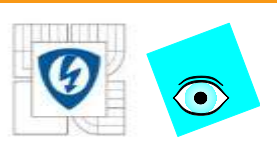
Průzkumová analýza dat (EDA)

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

Jednorozměrný výběr: sledován pouze jeden znak.

Počet prvků: (a) může být omezený a soubor je konečný,
(b) Může být neomezený a soubor je nekonečný.

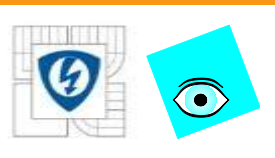
Náhodný výběr, $\{x_i\}, i = 1, \dots, n$, jehož prvky, tj. naměřené hodnoty, jsou chápány jako realizace jisté náhodné veličiny.



Reprezentativní náhodný výběr

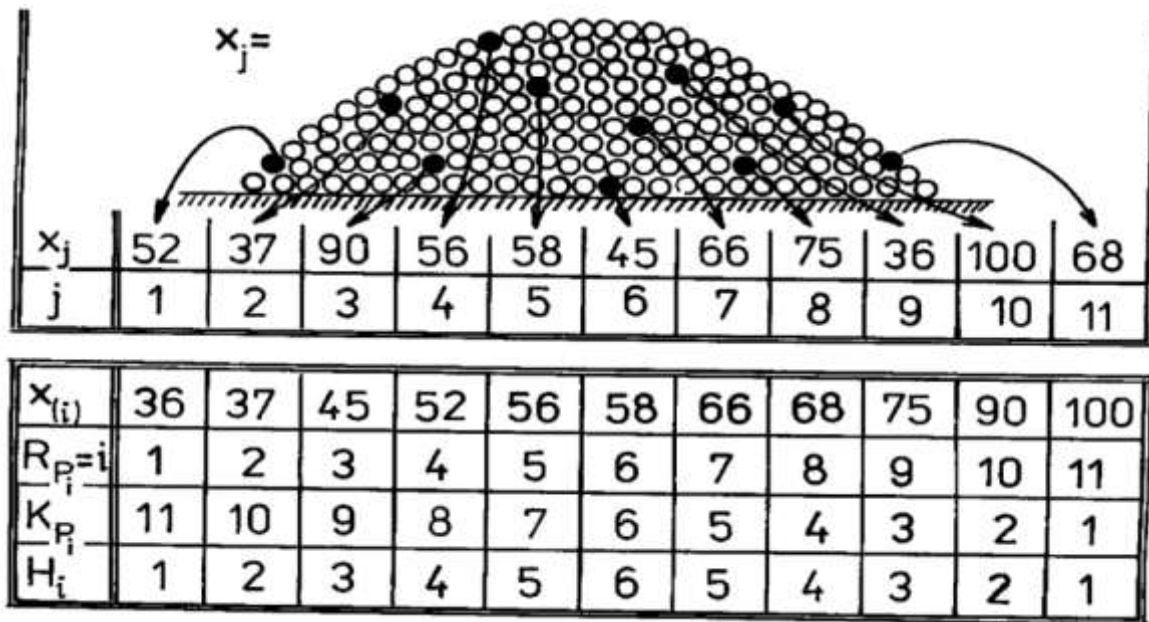
Je charakterizován předpoklady:

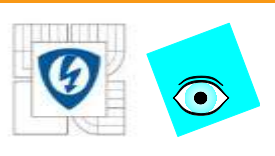
1. Jednotlivé prvky výběru x_i jsou **vzájemně nezávislé**.
2. Výběr je **homogenní**, tj. všechny jeho prvky x_i pocházejí ze stejného rozdělení s konstantním rozptylem, nejsou žádné vybočující prvky
3. Předpokládá se **normální** rozdělení pravděpodobnosti.
4. Všechny prvky souboru mají **stejnou pravděpodobnost**, že budou zařazeny do výběru.



Reprezentativní náhodný výběr

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





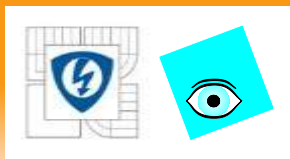
Technika pořadí a hloubek

Každá pořádková statistika x_j 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 $H_i = \min(R_{P_i}, K_{P_i})$.



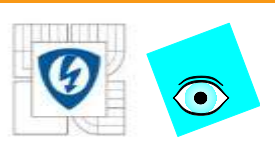
Pořádkové statistiky, pořadová pravděpodobnost

Pořádkové statistiky: vzestupně seříděné prvky výběru
 $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}.$

Pořadová pravděpodobnost je dána vztahem

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

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



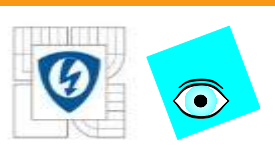
Rozptyl kvantilu

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

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

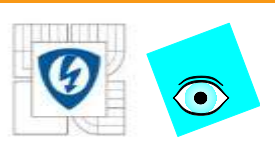
kde $f(\tilde{x}_\alpha)$ je výběrová hustota pravděpodobnosti v bodě $\tilde{x}_\alpha$.

Písmenové hodnoty: kvantily 'pro pořadové pravděpodobnosti $P_i = 2^{-i}, i = 1, 2, \dots,$



Rozptyl kvantilu

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,

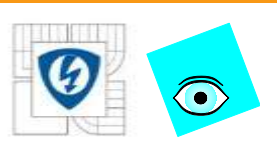
$L_H = F_H$, horní kvartil,

$i = 3, L_D = E_D$, dolní oktil,

$L_H = E_H$, horní oktil,

$i = 4, L_D = D_D$, dolní decil,

$L_H = D_H$, horní decil, ... atd.



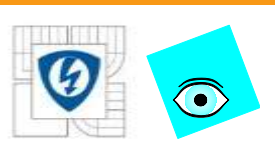
Hloubka dolních písmenných hodnot

Vypočte se dle

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

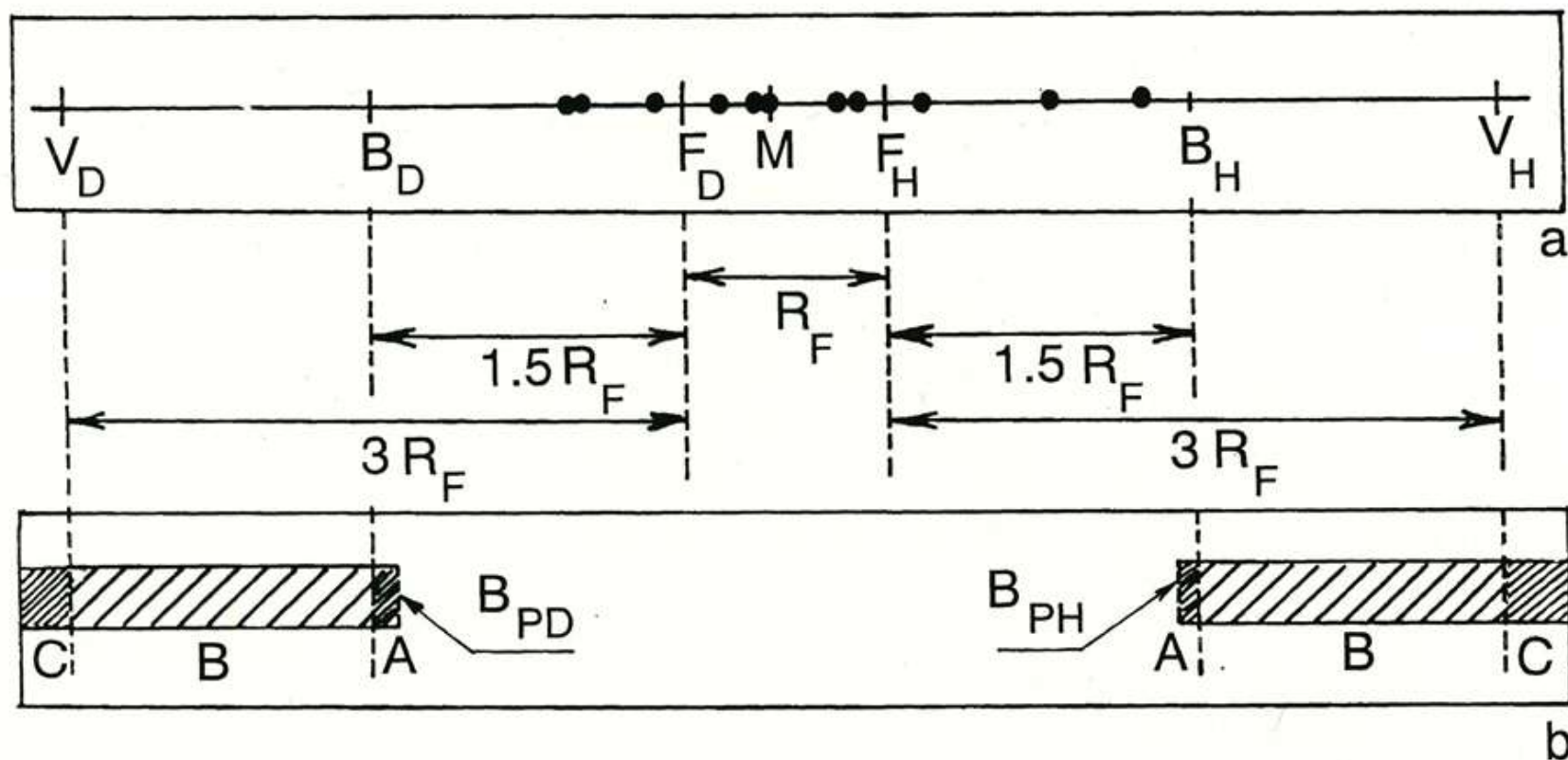
kde L jsou indexy F, E, D a $\text{int}(x)$ značí celočíselnou část čísla x .

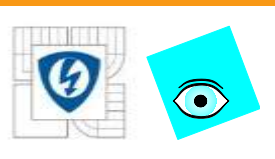
Pokud je $L = F$, bere se $L - 1 = M$.



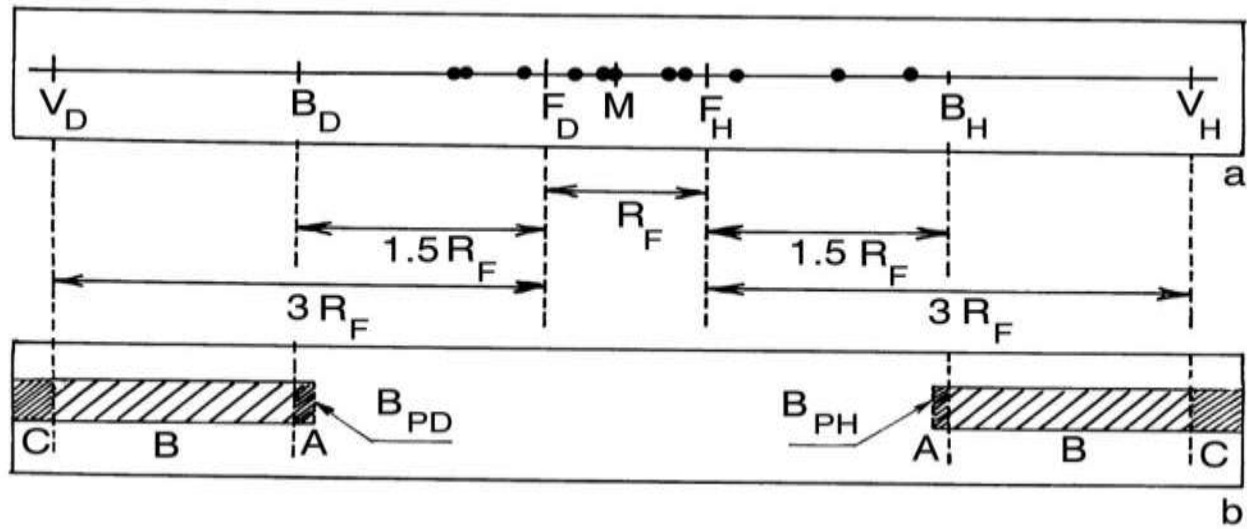
Užívání písmenových hodnot

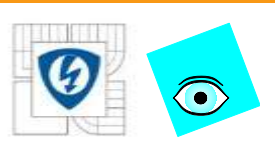
Kvartilové rozpětí: $R_F = F_H - F_D$





Užívání písmenových hodnot





Užívání písmenových hodnot

Přilehlé hodnoty:

$B_H = F_H + 1.5R_F$ horní přilehlá hodnota

$B_D = F_D - 1.5R_F$ dolní přilehlá hodnota

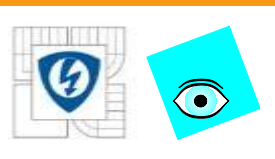
Vnitřní hradby:

$V_H = F_H + 3R_F$ horní vnitřní hradba

$V_D = F_D - 3R_F$ dolní vnitřní hradba

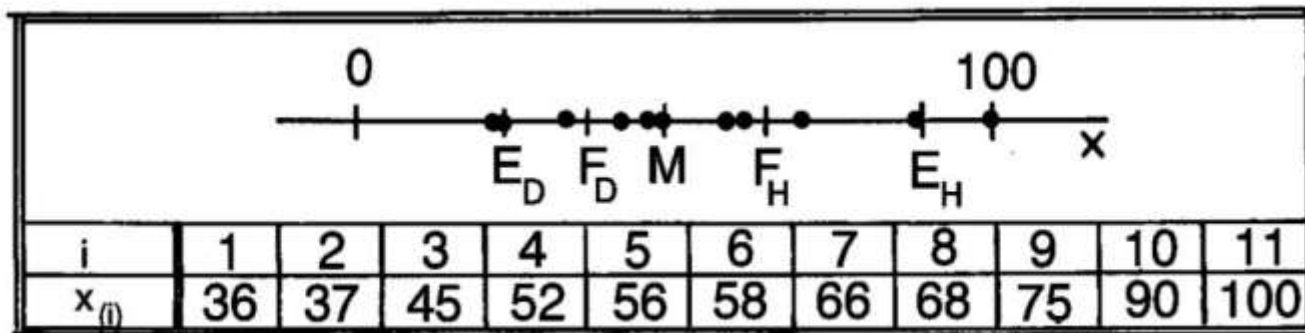
„Podezřelá měření“ (= vybočující hodnoty):

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



Graficko-tabelární schéma sumarizace dat

- a) obecné schéma písmenově-číslicového zápisu výběru,
b) sedmipísmenový zápis výběru.

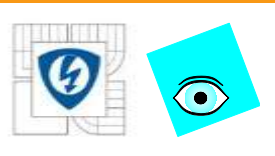


(a)

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

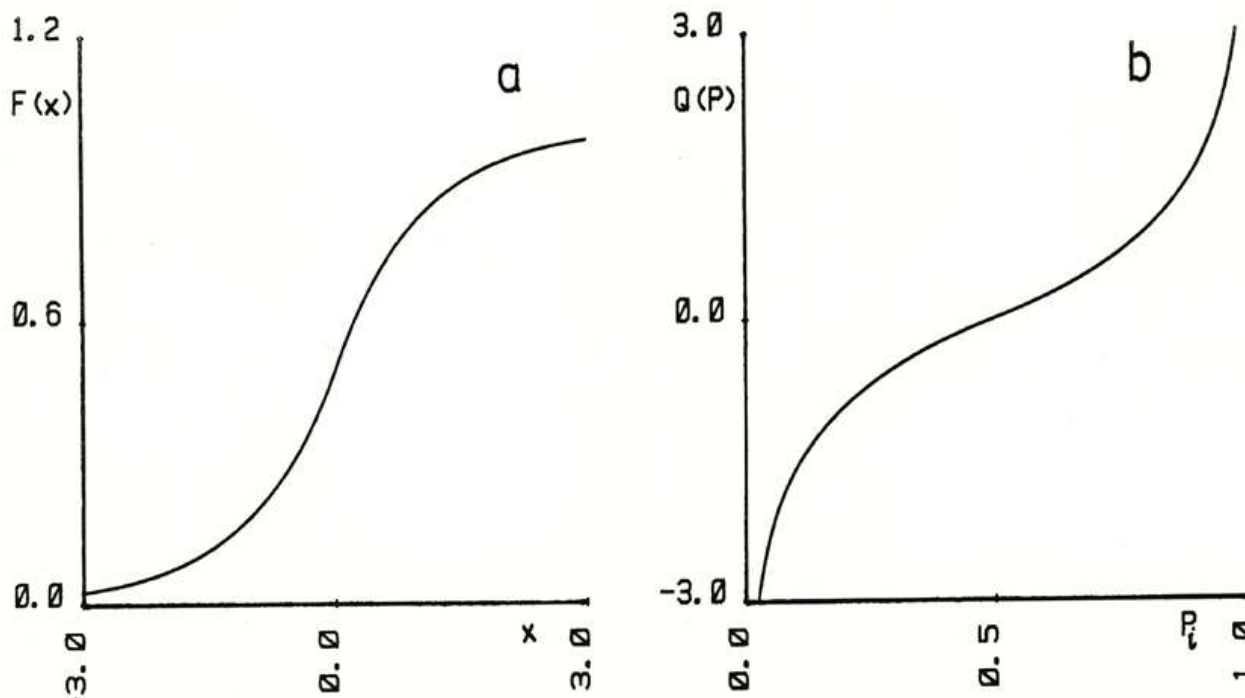
(b) $n=11$

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

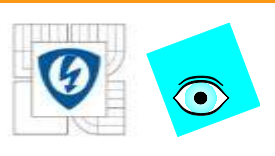


Kvantilová funkce

$Q(P)$: $x_{(i)}$ proti $P_i, i = 1, \dots, n$, 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



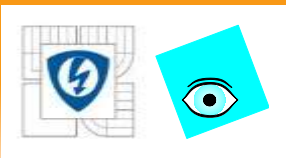
Střední hodnota i -té pořádkové statistiky

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.

$100\alpha\%$ ní kvantil $\tilde{x}_\alpha$: pro α z intervalu $\langle 0,1 \rangle$ 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

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

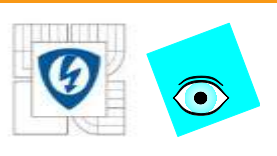


IDENTIFIKACE STATISTICKÝCH ZVLÁŠTNOSTÍ VÝBĚRU DAT

31.1.2011

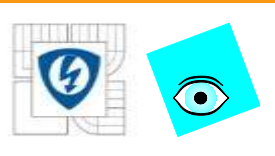
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





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:

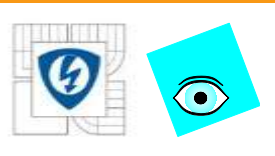
- G18 Graf poměrů frekvencí
- G19 Poissonův graf
- G20 Modifikovaný Poissonův graf

Spojité rozdělení (transformace):

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

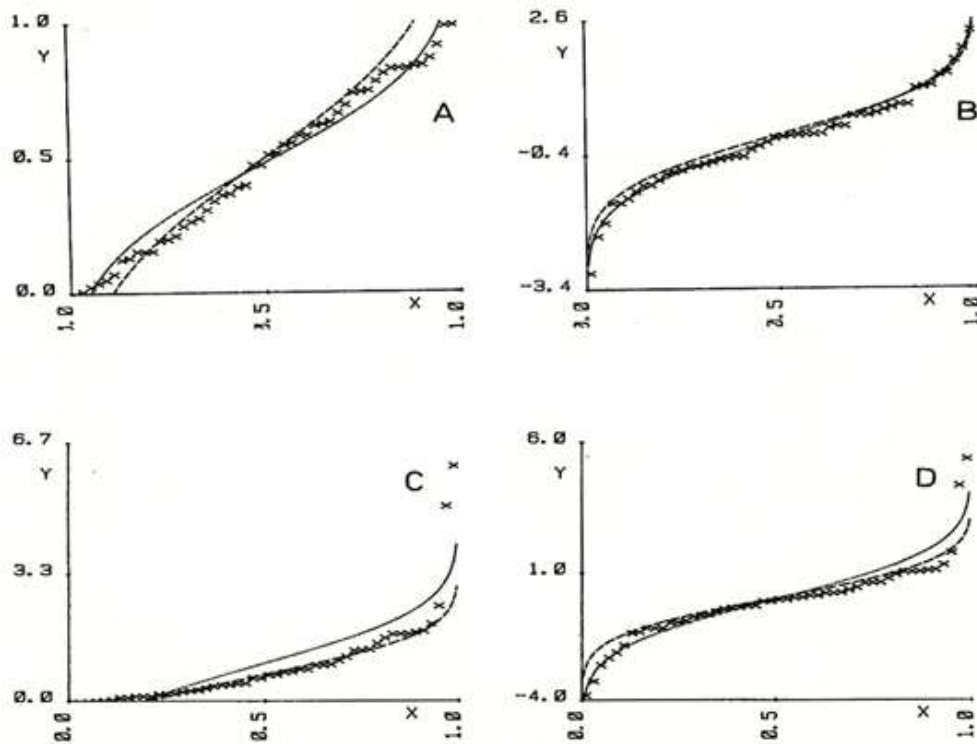
Testy:

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

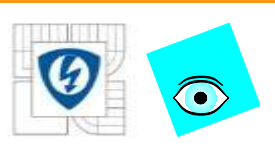


Kvantilový graf

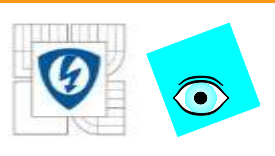
Osa x : pořadová pravděpodobnost P_i , Osa y : pořádková statistika $x_{(i)}$,



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



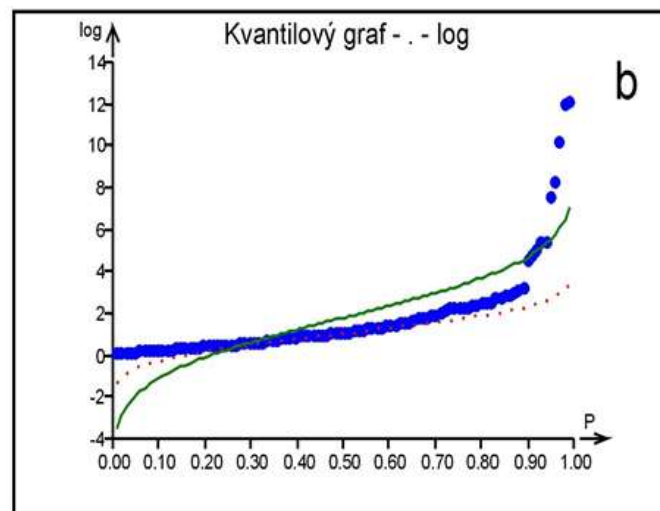
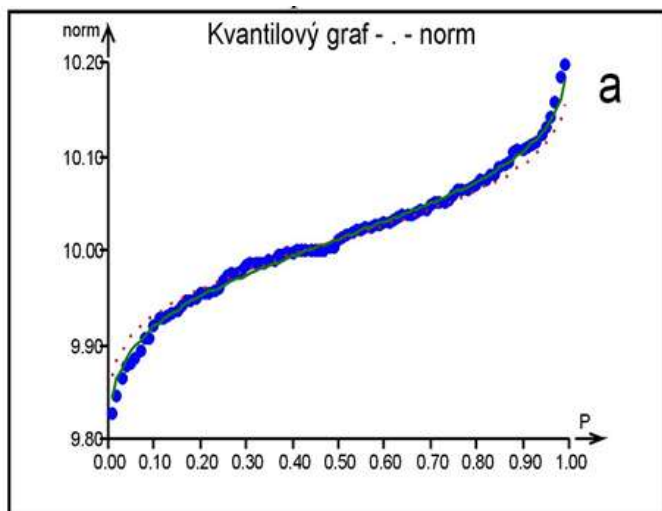
- Přehledně znázornit data, lokální koncentraci dat,
- Rozlišit tvar rozdělení (symetrický, sešikmený)
- Identifikovat vybočující data, atd.
- Zakreslují se i kvantilové funkce normálního rozdělení
- $N_{P_i} = \hat{\mu} + \hat{\sigma} \cdot u_{P_i}$ pro $0 \leq P_i \leq 1$:
- (.....) Klasických odhadů $\hat{\mu} = \bar{x}$ a $\hat{\sigma} = s$,
- (----) Robustních odhadů $\hat{\mu} = \tilde{x}_{0.5}$ a $\hat{\sigma} = \frac{R_F}{1.349}$,



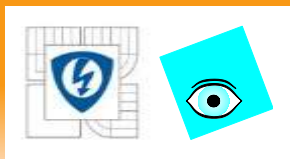
Osa x : pořadová pravděpodobnost P_i , Osa y : pořádková statistika $x_{(i)}$,

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

- (1) Klasických odhadů parametrů polohy a rozptýlení, tj. aritmetického průměru a směrodatné odchylky $\hat{\mu} = \bar{x}$ 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 interkvantilové rozpětí.



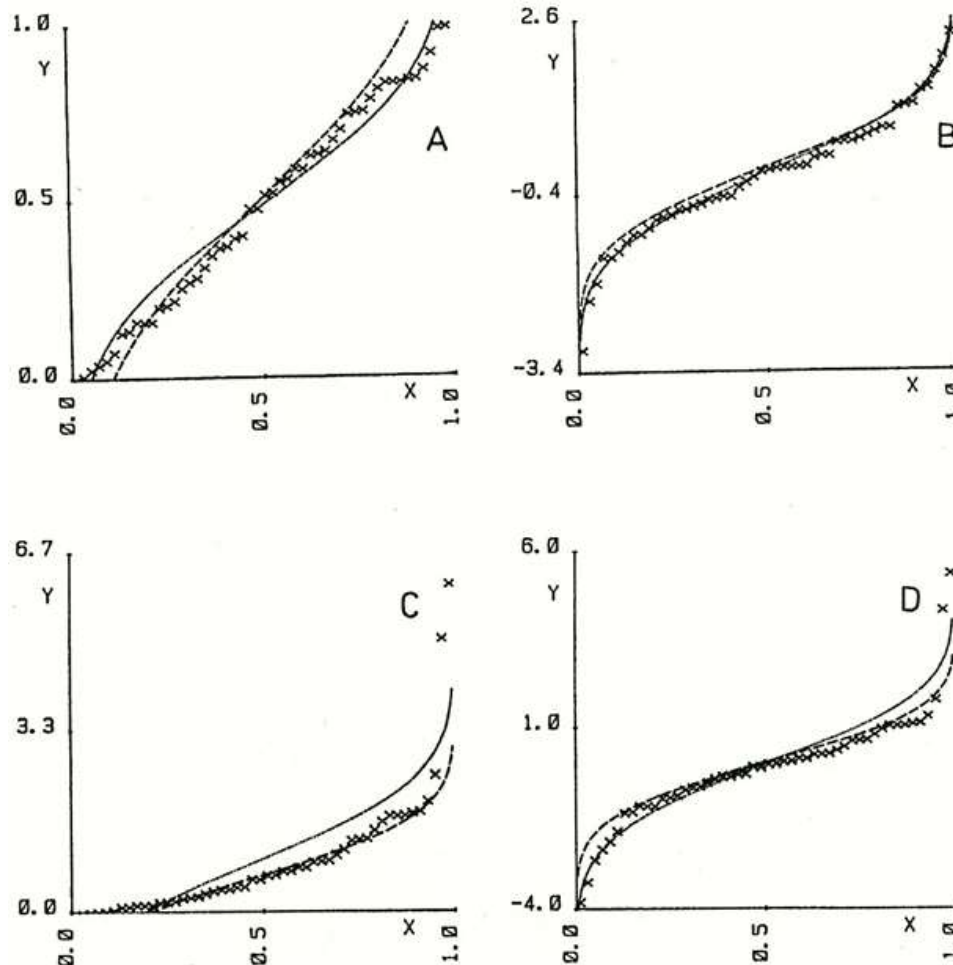
Kvantilový graf (robustní --- a klasický ...) pro výběry: (A) norm. symetrického (Gaussova, normálního), (B) log, symetrického (logaritmicko-normálního) rozdělení ADSTAT



Kvantilové grafy (robustní --- a klasické ...) pro výběry z rozdělení

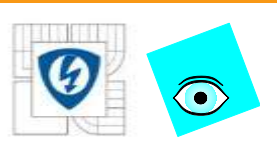
G1

(A) Rovnoměrné, (B) Normální, (C) Exponenciální, (D) Laplaceovo



31.1.2011

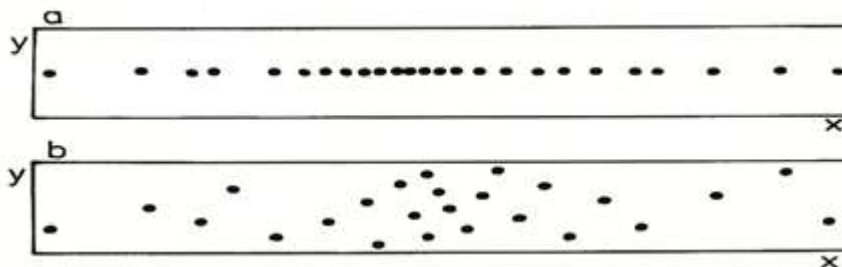
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



Osa x : hodnoty x

Osa y : libovolná úroveň, obvykle $y = 0$

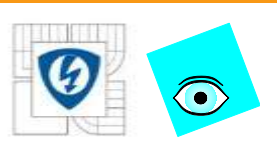
Obsah: jednorozměrná projekce kvantilového grafu do osy x .



Konstrukce (a) diagramu rozptýlení a (b) rozmítnutého diagramu rozptýlení.

Diagnóza:

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



Osa x : hodnoty x

Osa y : libovolná úroveň, obvykle $y = 0$

Obsah: představuje rozmítanou projekci kvantilového grafu

Diagnóza: přehledně znázornit data, lokální koncentraci dat, indikuje i podezřelá a vybočující měření.

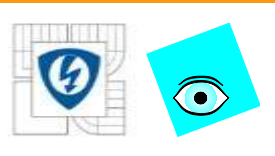


Diagram p rozptýlení

Osa x : hodnoty x

Osa y : libovolná úroveň, například $y = 0$

Představuje jednorozměrnou projekci kvantilového grafu do osy x . I při své jednoduchosti ukazuje na lokální koncentraci dat a indikuje podezřelá a odlehlá měření.

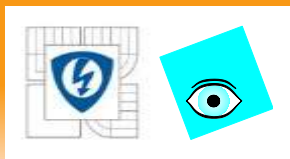
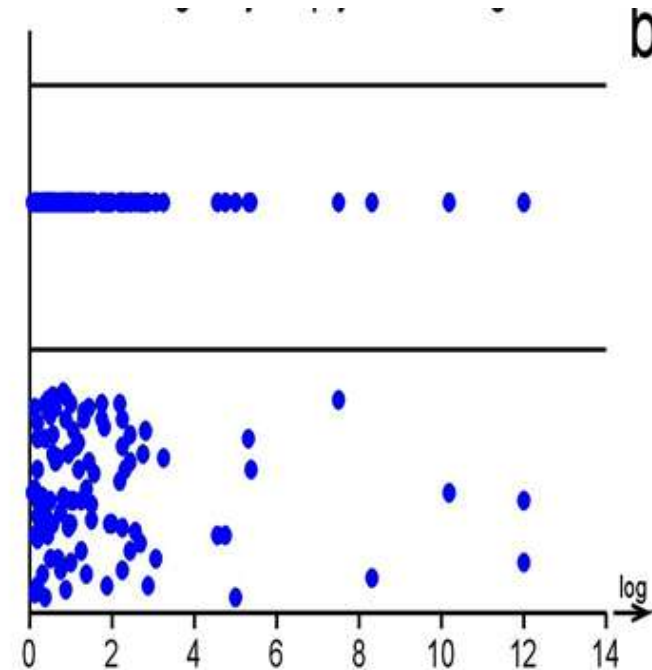
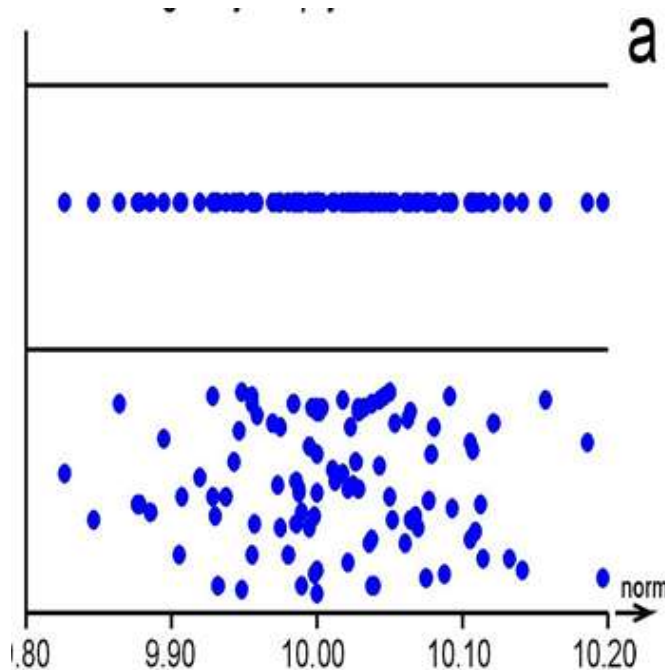


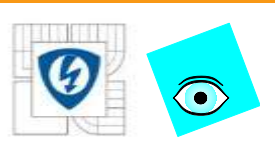
Diagram rozptýlení a rozmítnutý diagram rozptýlení

Pro výběry (shora dolů) (a) *norm*, symetrického (normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ADSTAT.



G2

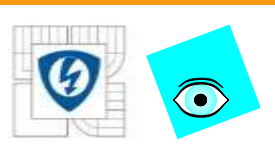
G3



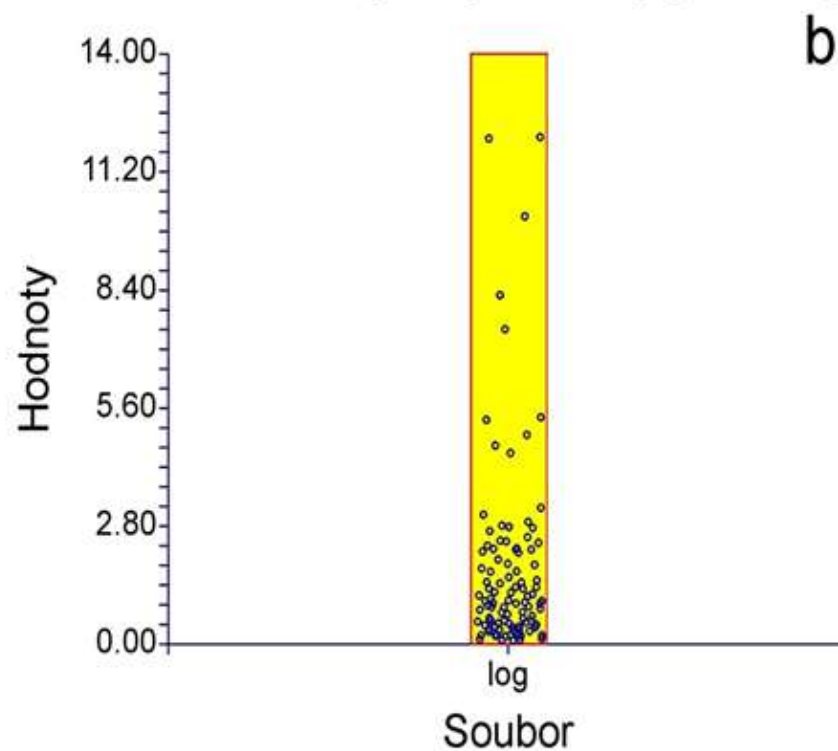
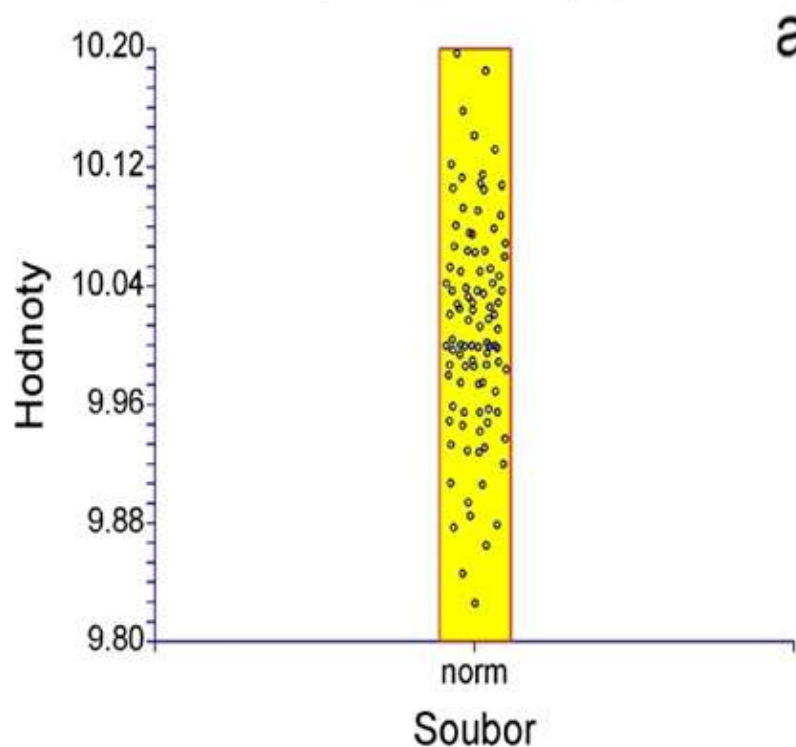
Osa x : hodnoty x

Osa y : interval náhodných čísel

Diagram představuje rovněž projekci kvantilového grafu, body jsou však pro lepší přehlednost vhodně rozmítnuté.



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



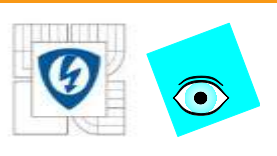


Diagram percentilů

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.

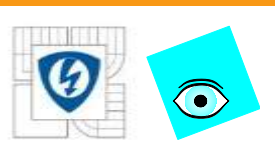
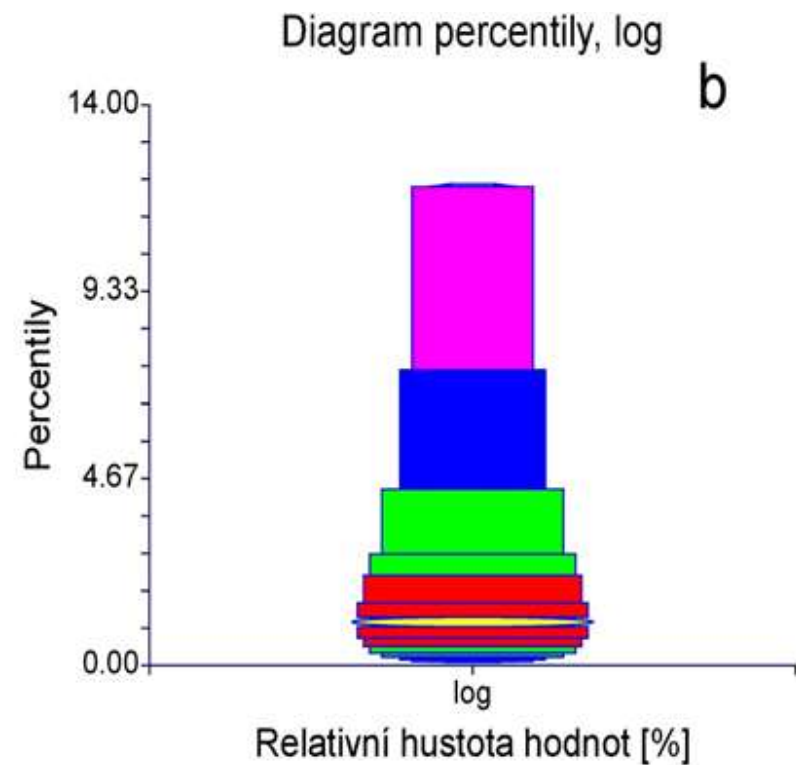
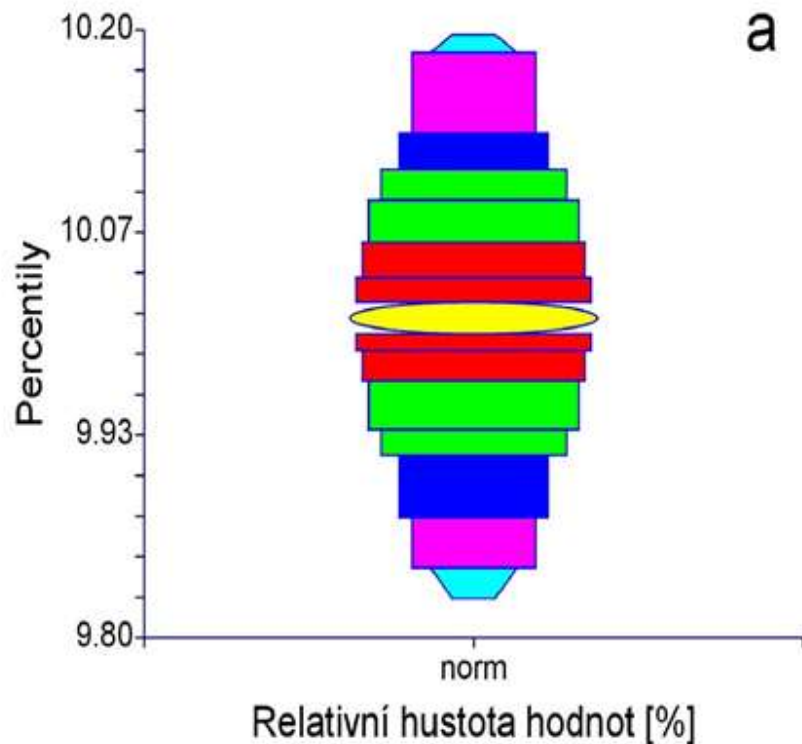
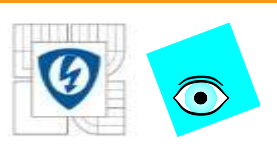


Diagram některých percentilů pro výběry

(a) *norm*, symetrického (Gaussova), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, NCSS2000



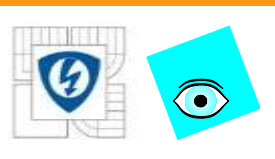


Houslový diagram

Osa x : název výběru proměnné

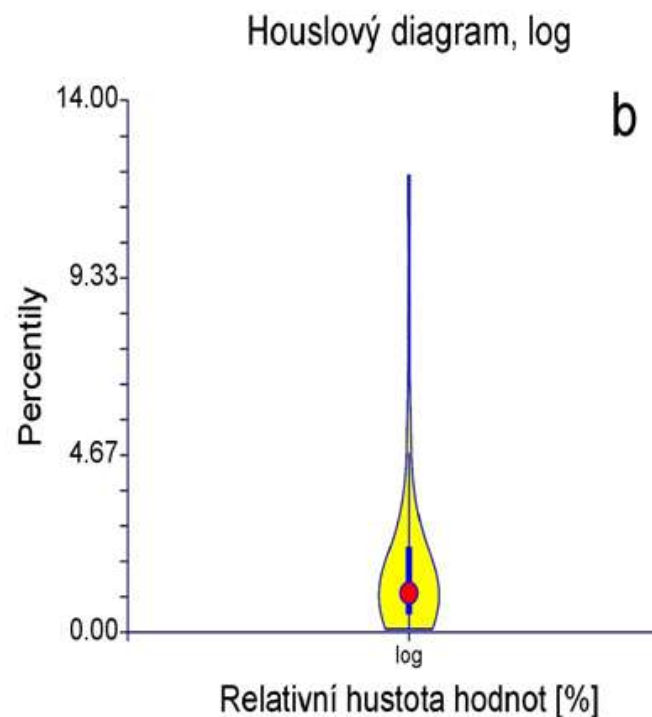
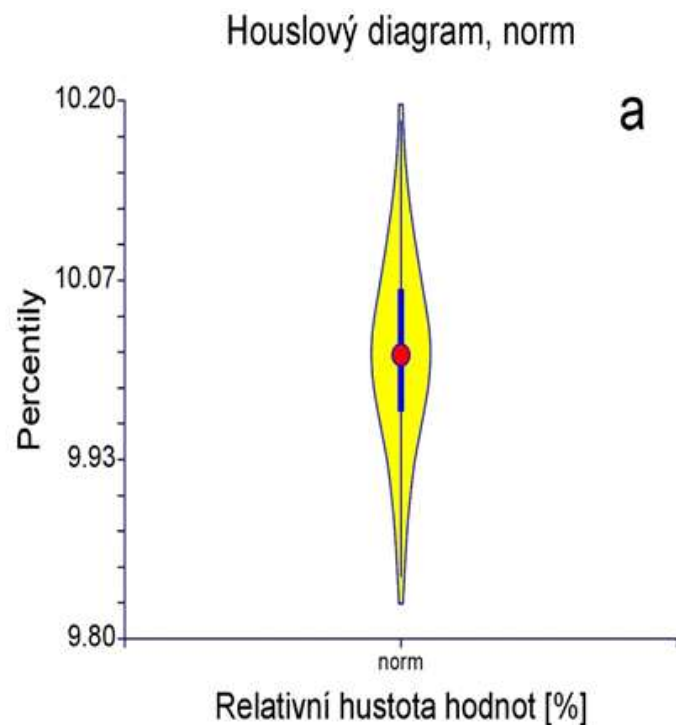
Osa y : percentily, hodnoty proměnné

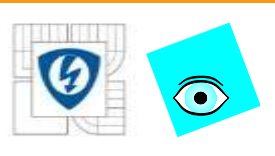
Medián je zobrazen černým kolečkem a začátek a konec úsečky zobrazuje horní a dolní kvantil. Normální rozdělení se projeví v symetrickém tvaru houslí, zatímco logaritmicko normální v silně asymetrickém tvaru.



Houslový diagram pro výběry

(a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, NCSS2000



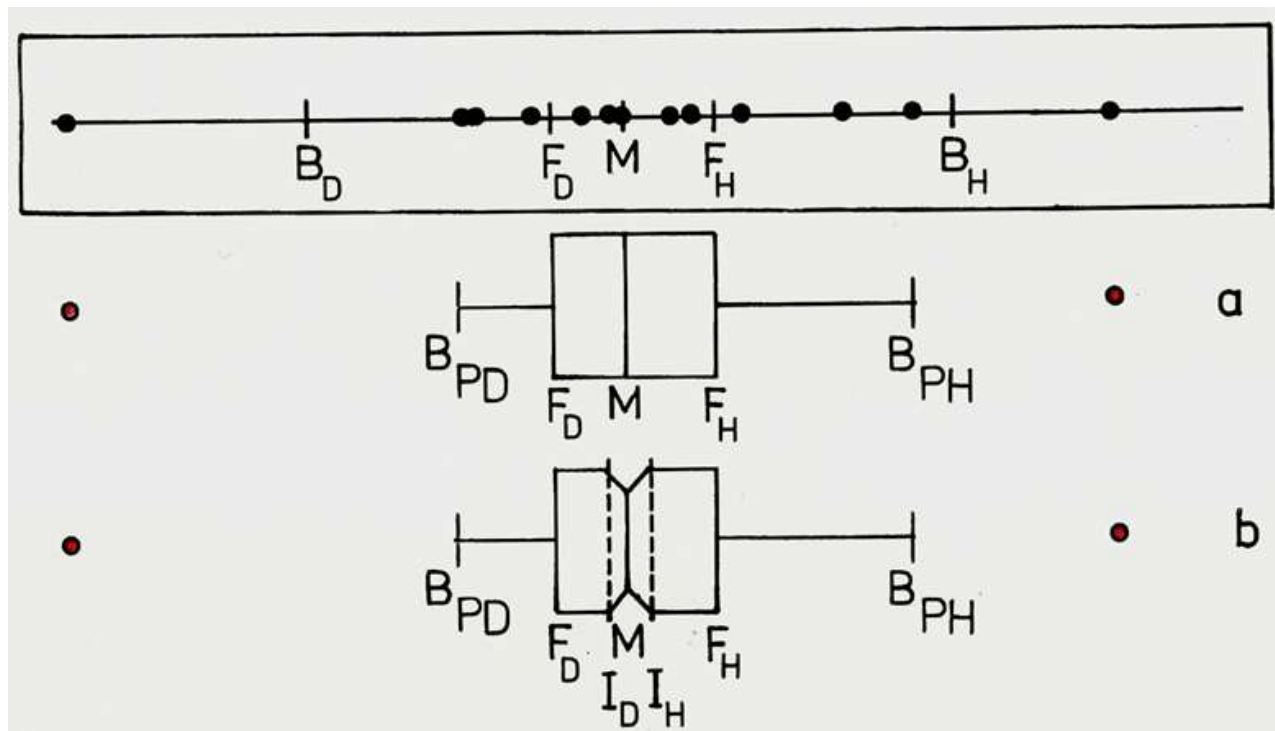


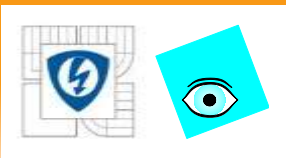
Krabicový graf

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.





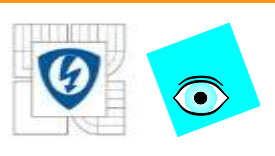
Krabicový graf

G4

Osa x : úměrná hodnotám x , Osa y : interval úměrný $\sqrt{n}$

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

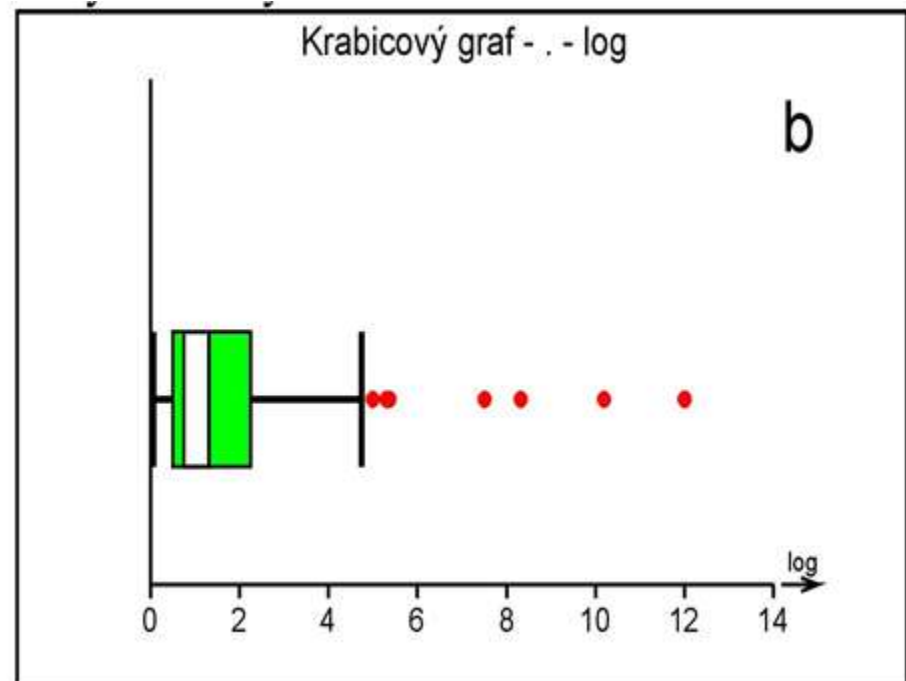
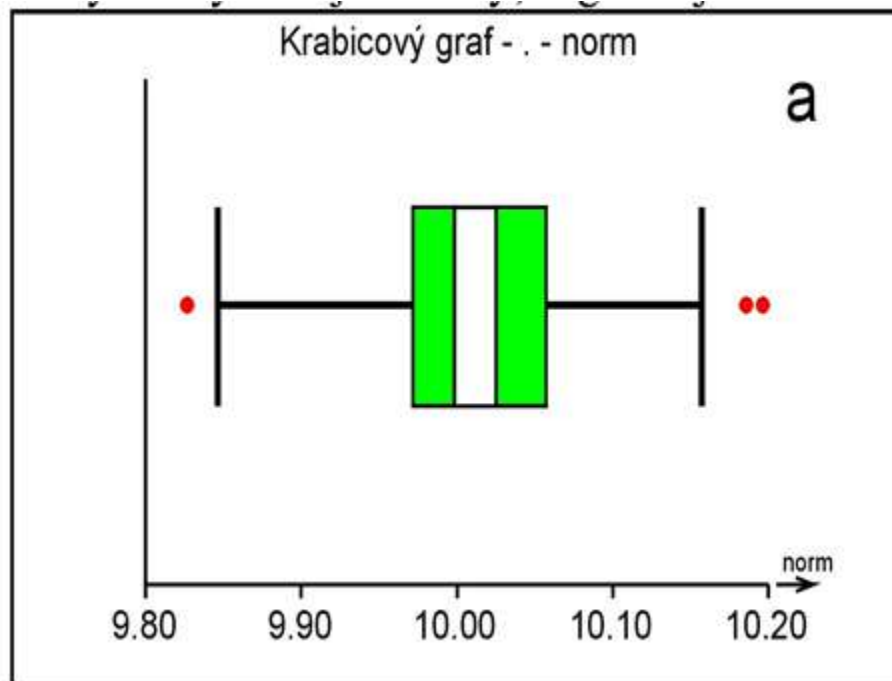
Prvky výběru, ležící mimo interval vnitřních hradeb $[B_H, B_D]$ jsou považovány za podezřelé, obvykle vybočující body; v grafu jsou znázorněny kroužky.



Vrubový krabicový graf pro výběry

G4

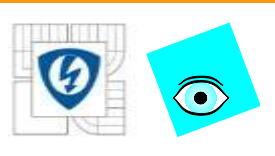
(a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ADSTAT.



31.1.2011

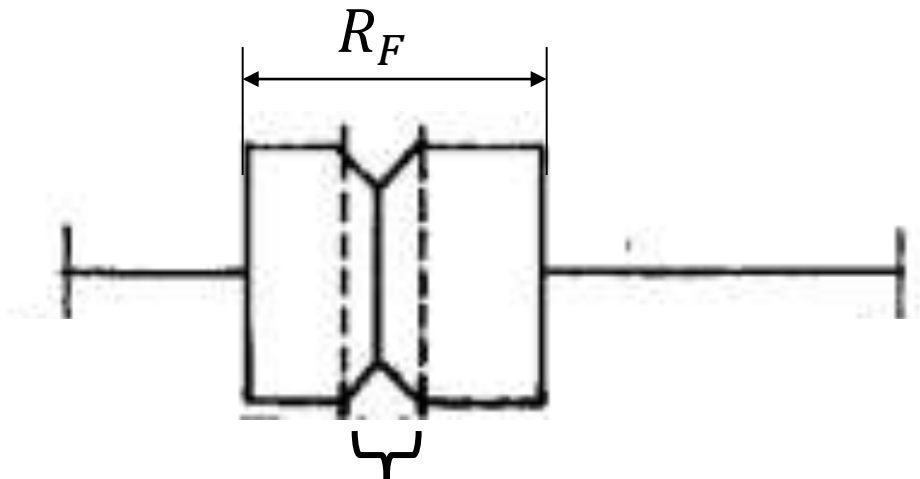
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

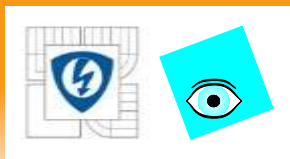




Intervalový odhad mediánu

$$I_{D,H} = M \pm \frac{1.57 \cdot R_F}{\sqrt{n}}$$





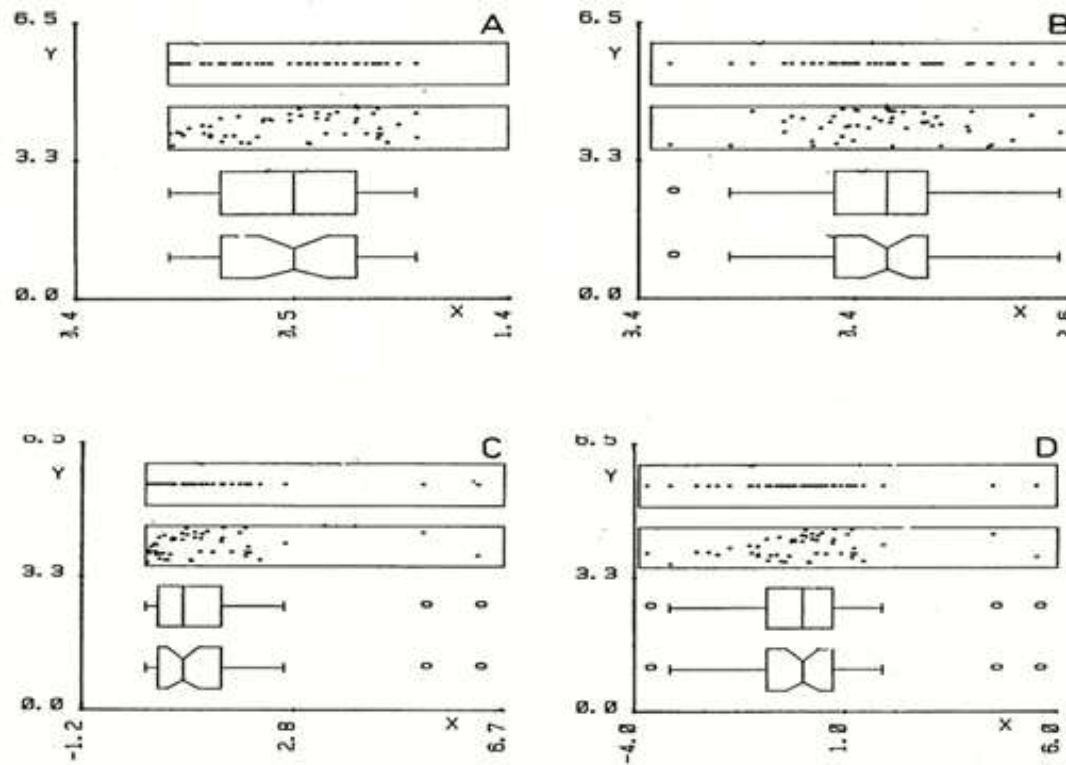
Vrubový krabicový graf

G5

Osa x : úměrná hodnotám x , Osa y : libovolný interval

Diagramy rozptýlení a krabicové grafy výběrů rozdělení

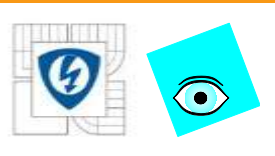
(A) Rovnoměrného, (B) Normálního, (C) Exponenciálního a (D) Laplaceova



31.1.2011

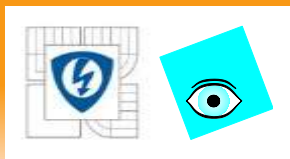
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





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

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



Krabicový vrubový krabicový graf

Krabicový graf: obsahuje medián, kvartily, přilehlé hodnoty a vnitřní hradby.

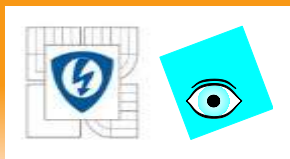
G4

Vrubový krabicový graf: Obsahuje navíc i interval spolehlivosti mediánu.

G5

Diagnóza:

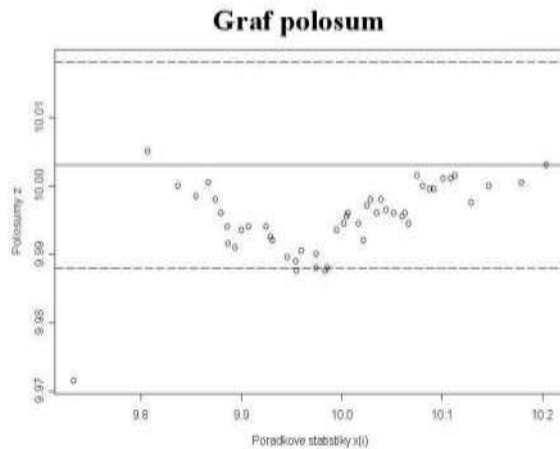
1. Robustní odhad polohy (medián)
2. Interval spolehlivosti mediánu
3. Posouzení symetrie kvantilů
4. Posouzení symetrie v okolí konců rozdělení
5. Identifikace odlehlých bodů, podezřelých bodů



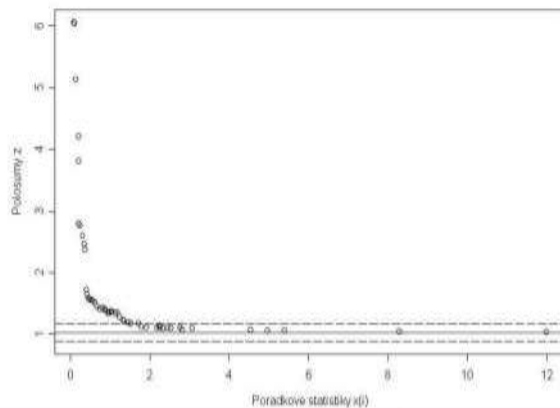
Graf polosum

G6

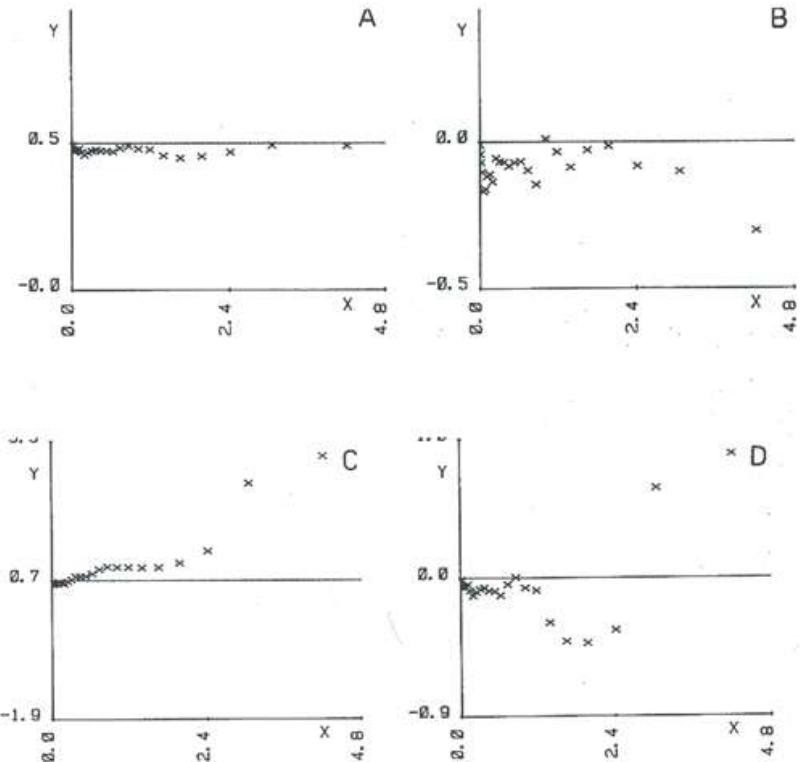
Osa x : pořádkové statistiky $x_{(i)}$, Osa y : $Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$



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



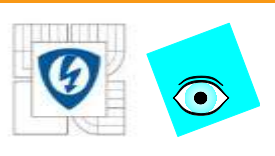
Graf polosum logaritmicke-normálního rozdělení *log LN* $(5, 2)$



Grafy symetrie pro výběry rozdělení

(A) Rovnoměrného, (B) Normálního, (C) Exponenciálního
(B) a (D) Laplaceova, (... značí symetrii)

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

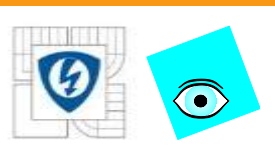


Graf polosum

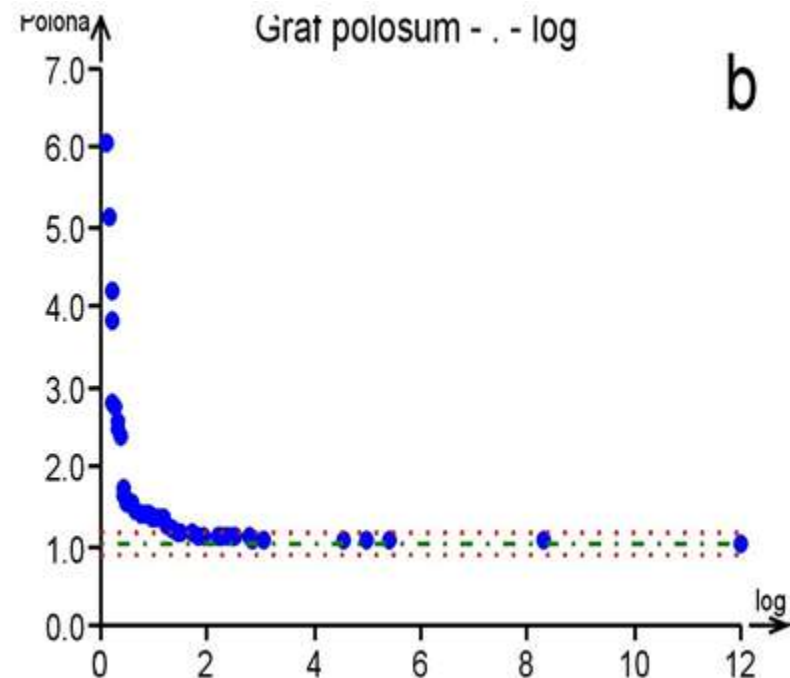
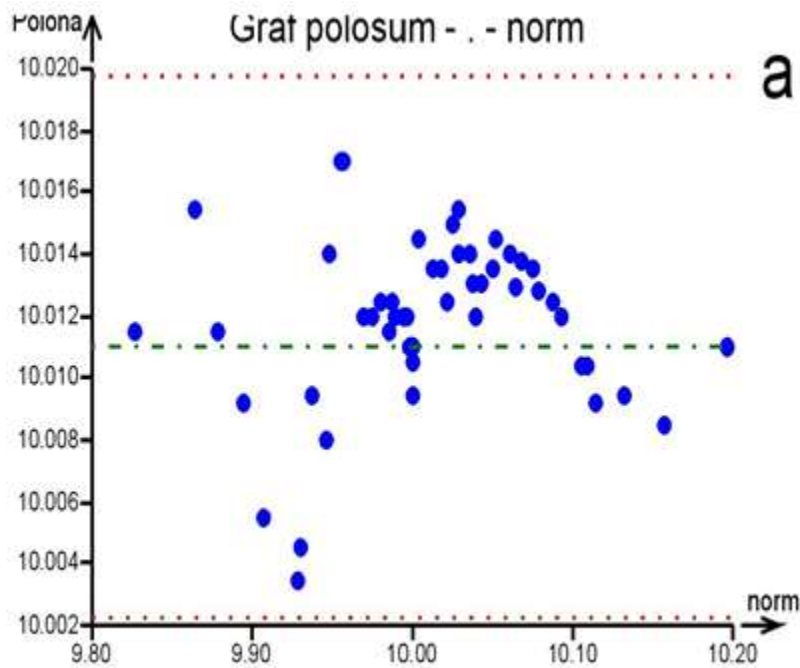
G6

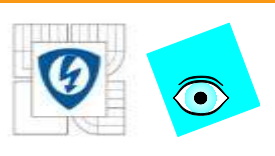
Osa x : pořádkové statistiky $x_{(i)}$, Osa y : $Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$

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.



(a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ADSTAT.



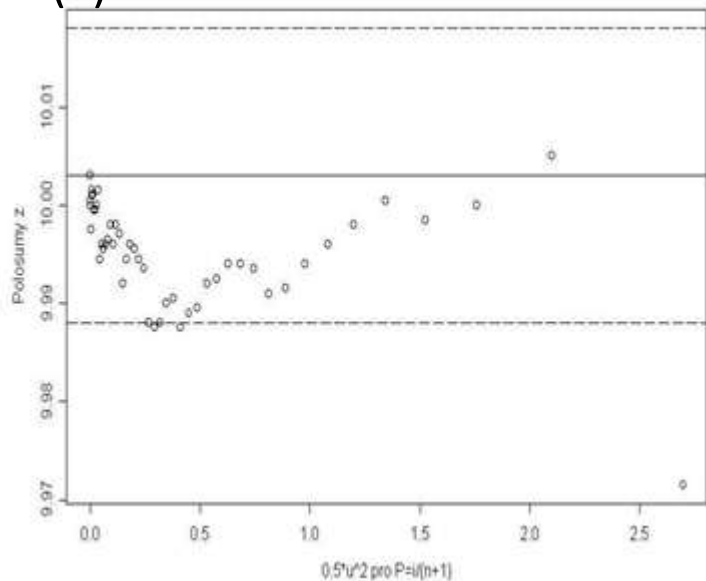


Graf symetrie

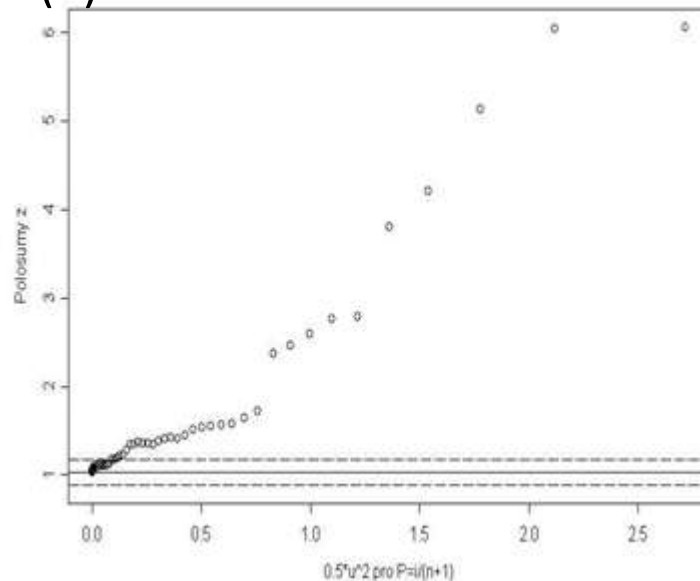
(a) Graf symetrie Gaussova normálního rozdělení
 $N(10,0.1)$

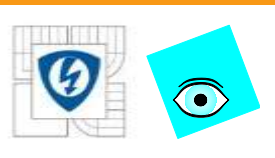
(b) Graf symetrie logaritmicko normálního rozdělení
 $\log LN(5,2)$

(a)



(b)

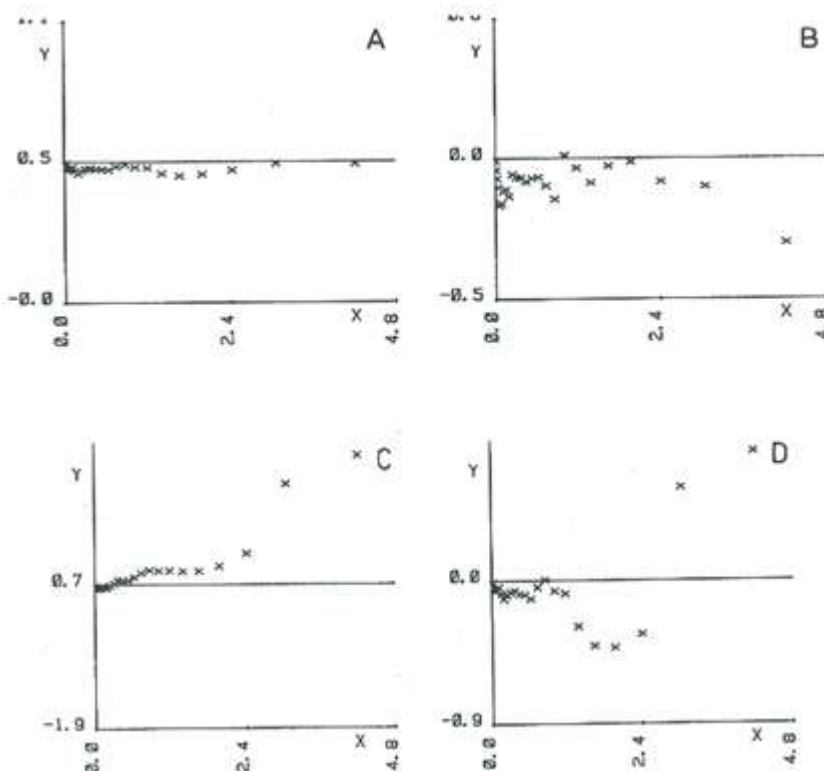




Grafy symetrie pro výběry rozdělení

Rovnoměrného, (B) Normálního, (C) Exponenciálního a (D) Laplaceova, (... značí symetrii)

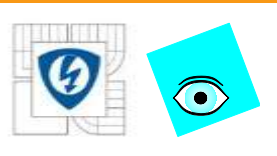
Osa x : $\frac{u_{P_i}^2}{2}$ pro y $P_i = \frac{i}{n+1}$, Osa y : $Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$



Diagnóza:

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

Směrnice je odhadem šikmosti.

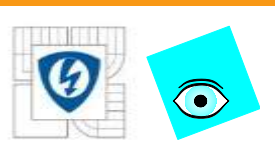


Graf symetrie

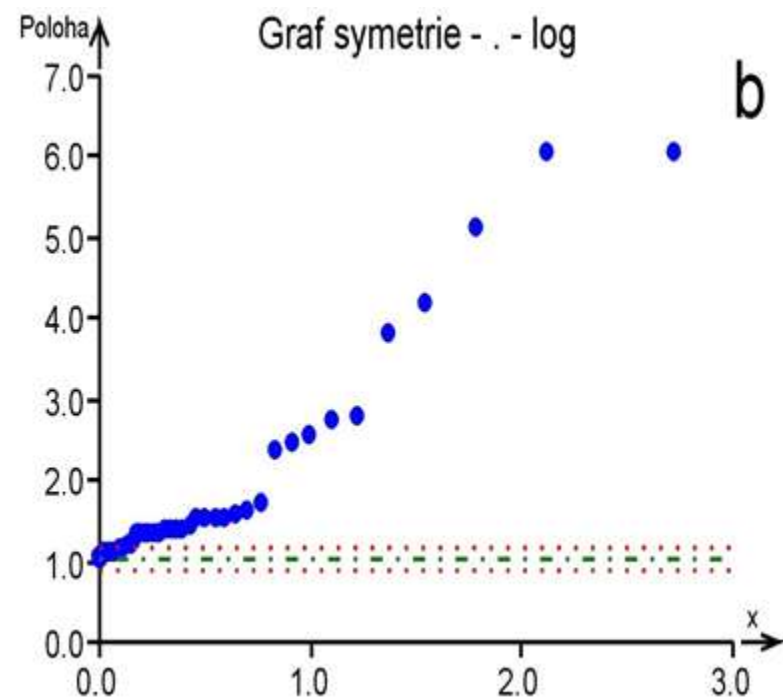
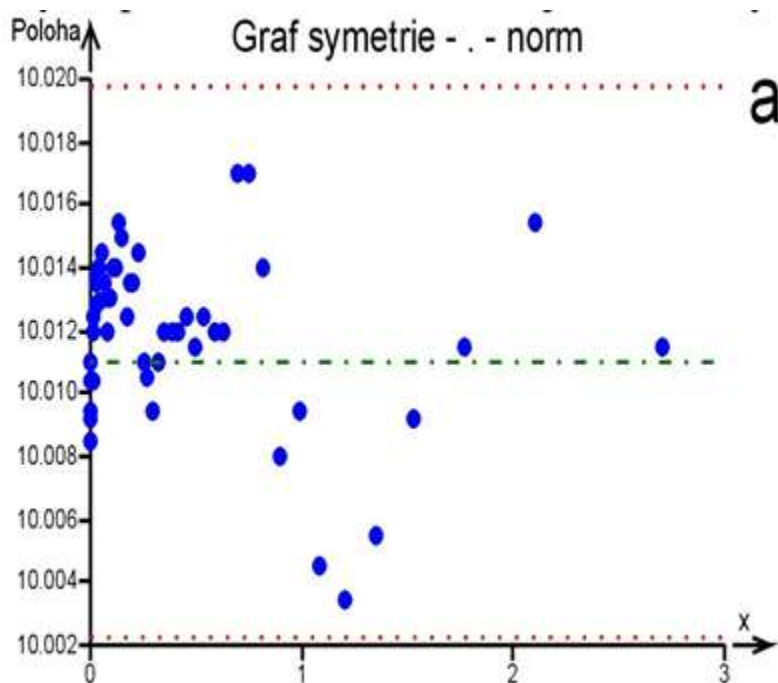
G7

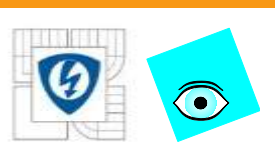
Osa x : $M - x_{(i)}$, Osa y : $x_{(n+1-i)} - M$

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.

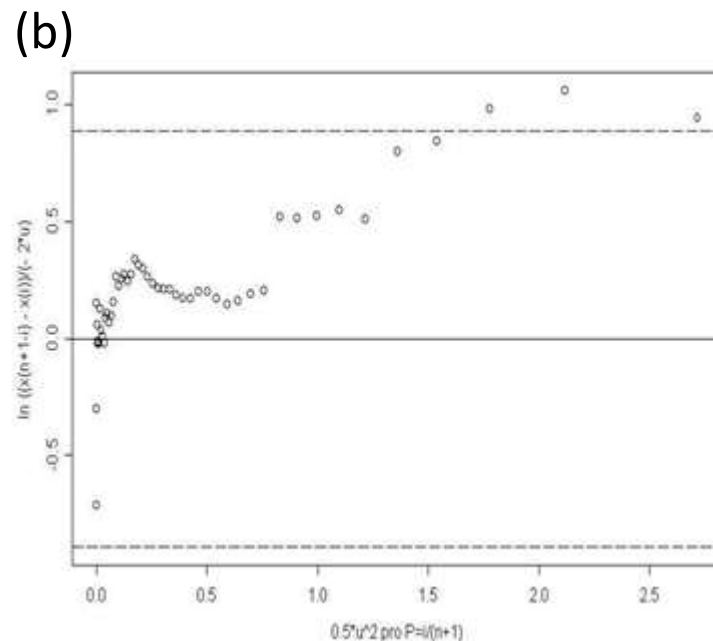
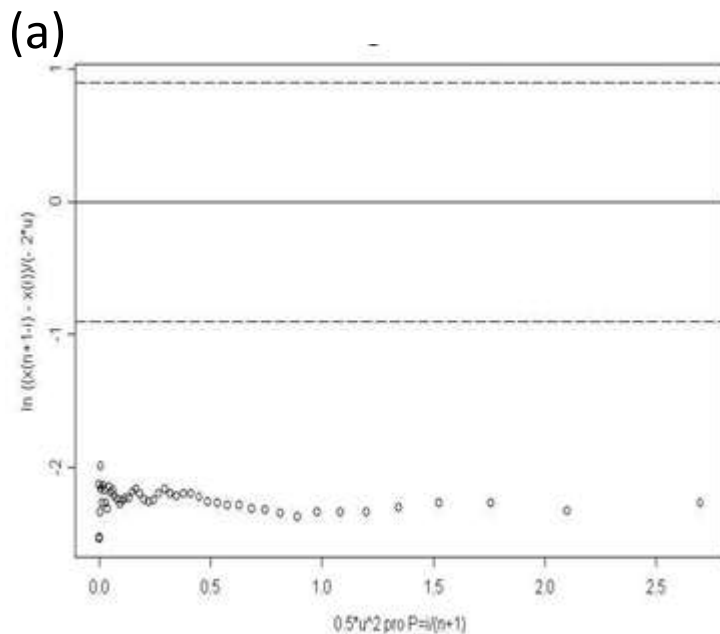


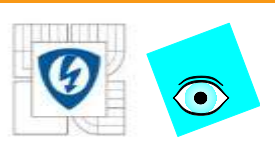
(a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ADSTAT.



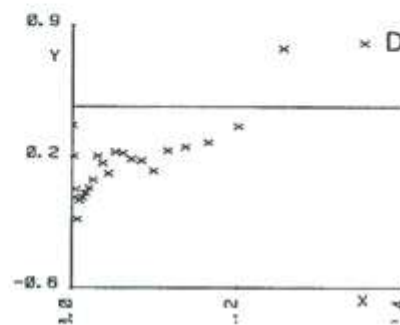
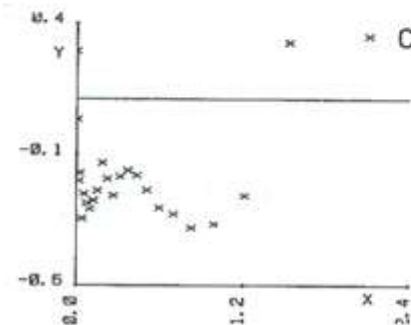
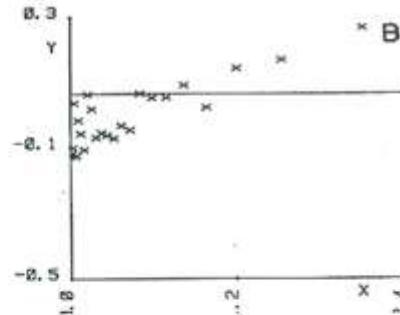
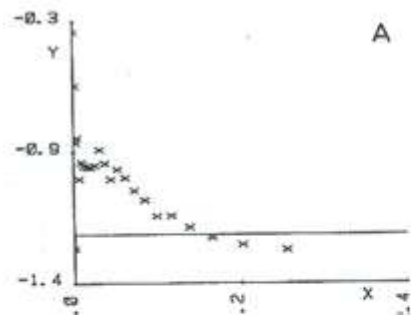


- (a) Graf špičatosti Gaussova normálního rozdělení $N(10,0.1)$
- (b) Graf špičatosti logaritmicko normálního rozdělení $\log LN(5,2)$



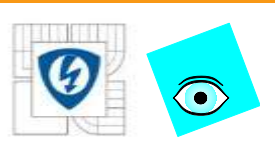


(A) Rovnoměrného, (B) Normálního, (C) Exponenciálního a (D) Laplaceova, (____) značí normalitu



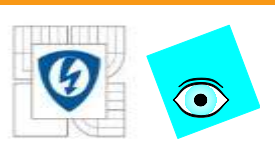
Diagnóza:

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

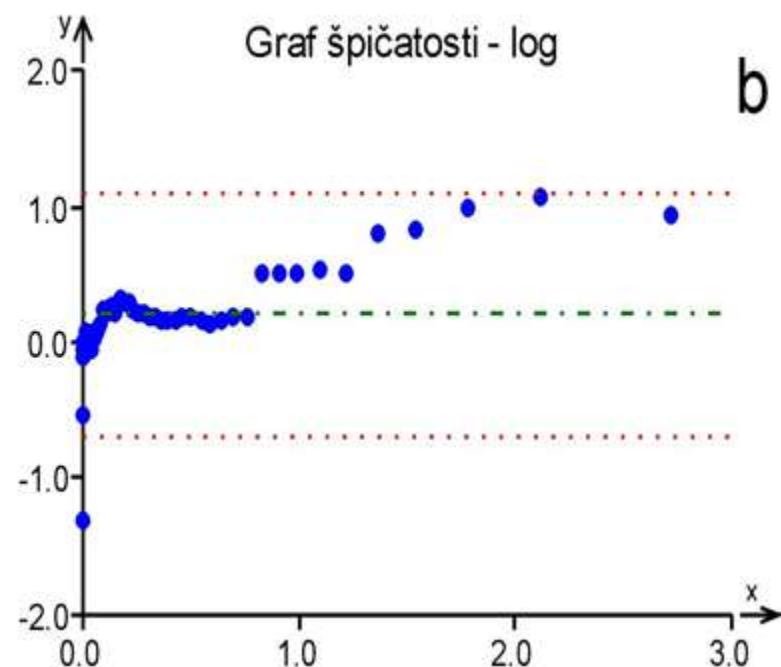
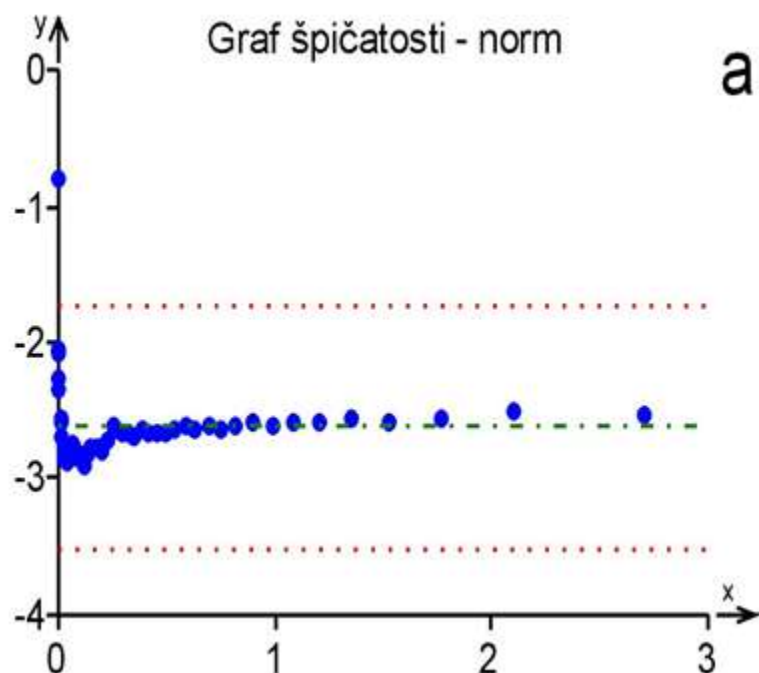


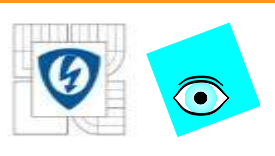
Osa x : $\frac{u_{P_i}^2}{2}$ pro y $P_i = \frac{i}{n+1}$, Osa y : $\ln(x_{(n+1-i)} / -2u_{P_i})$

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.



(a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ADSTAT.

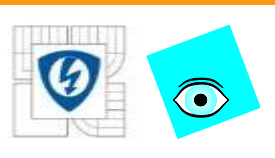




Graf šikmosti

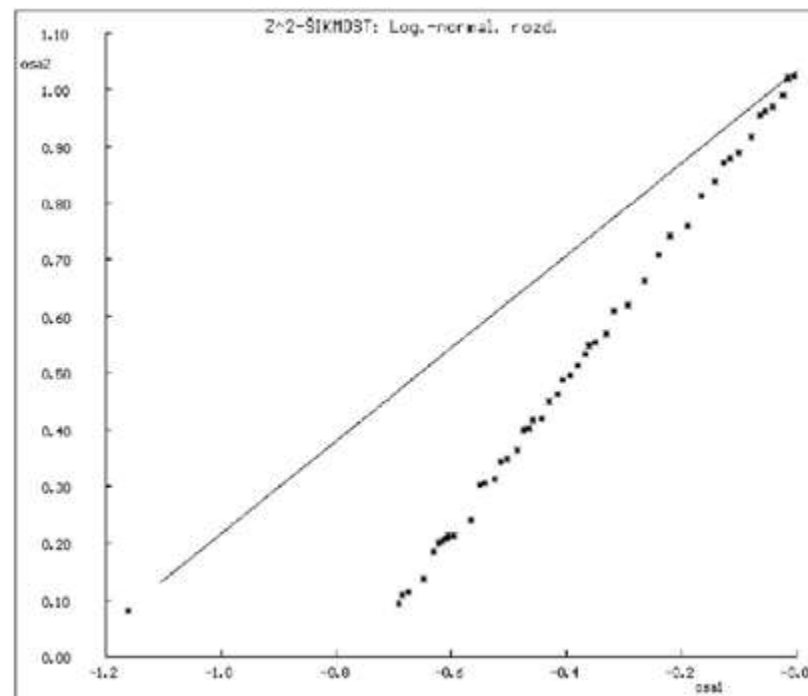
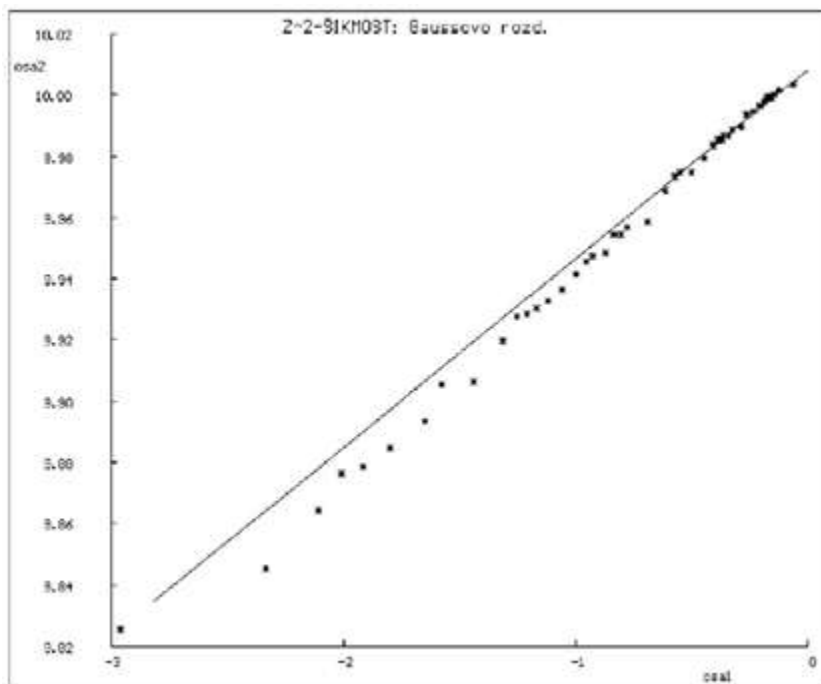
$$\text{Osa } x: \frac{u_{P_i}^2}{2} \text{ pro } y \ P_i = \frac{i}{n+1}, \text{ Osa}$$
$$y: Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$$

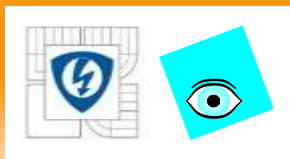
Pro případ symetrického rozdělení rezultuje u grafu šikmosti přímková závislost s nulovým úsekem a jednotkovou směrnicí. Body leží těsně na této přímce a vykazují jinou směrnici.



Graf šikmosti pro výběry

(a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ADSTAT.



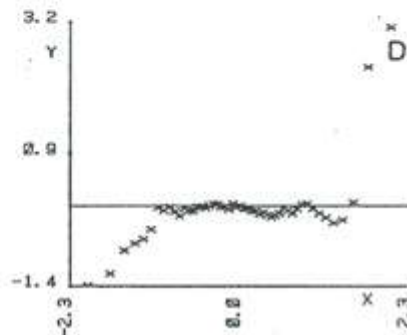
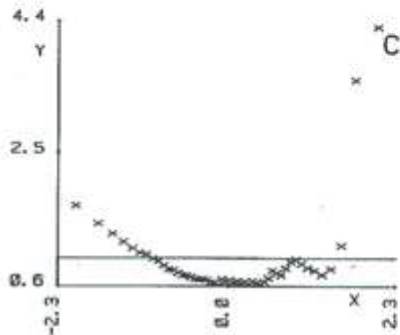
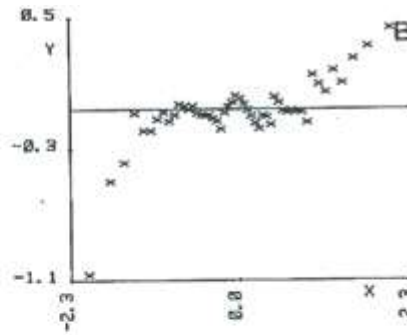
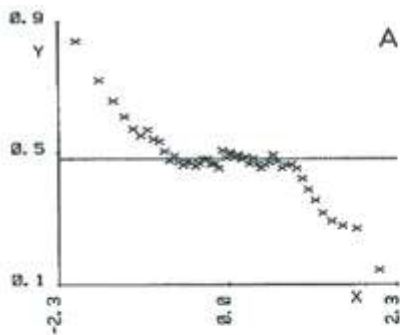


Diferenční kvantilový graf

G9

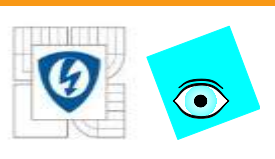
Osa x : kvantil u_{p_i} , Osa y : $d_{(i)} = x_{(i)} - \tilde{s}u_{p_i}$

Diferenční kvantilové grafy pro výběry rozdělání (A) Rovnoměrného, (B) Normálního, (C) Exponenciálního a (D) Laplaceova, (____) značí normalitu



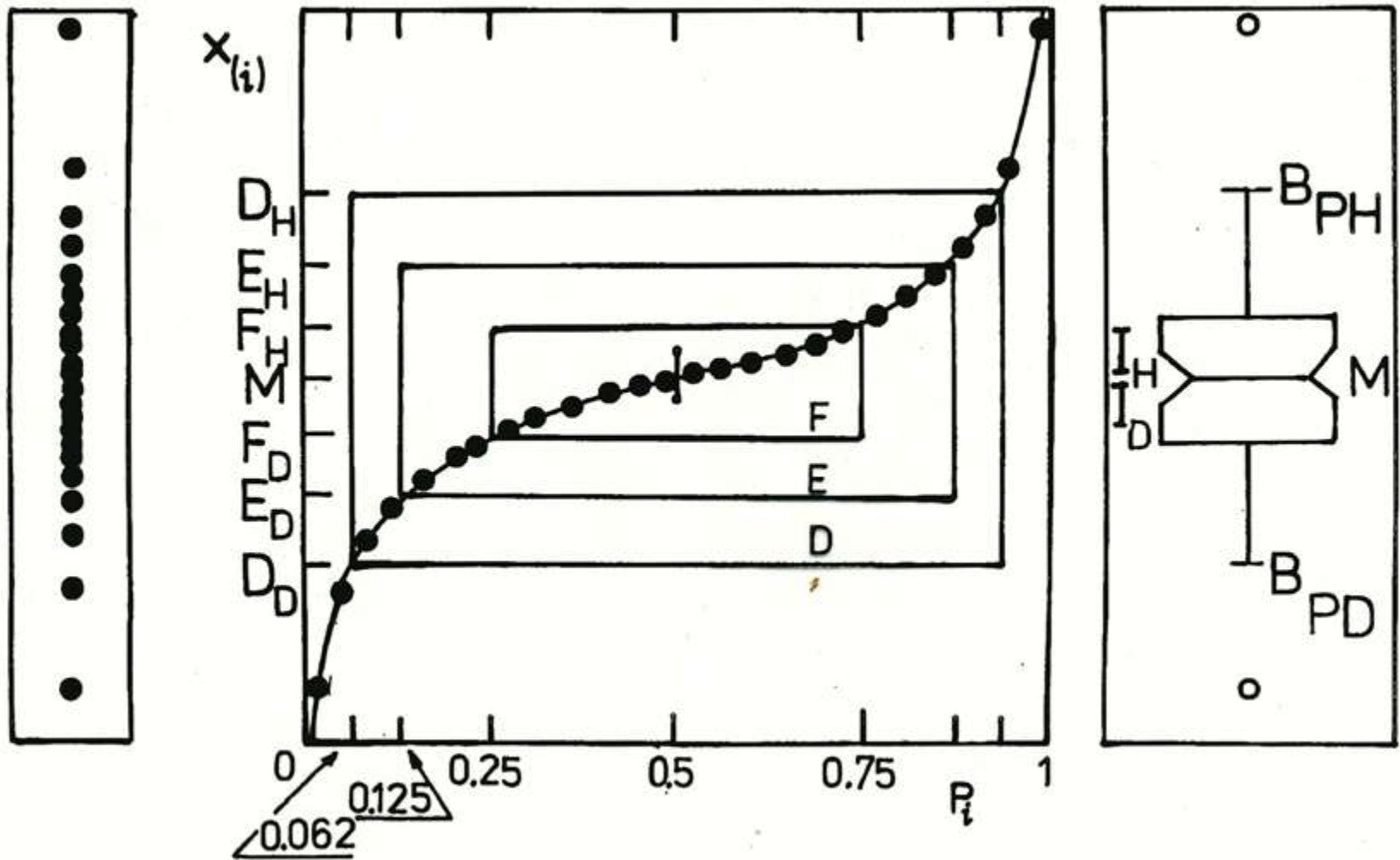
Diagnóza:

- (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

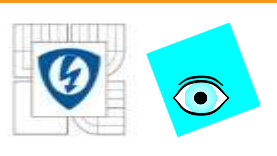
Osa x : P_i , Osa y : $x_{(i)}$



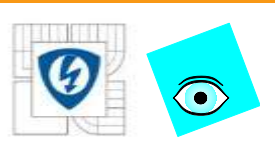
31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

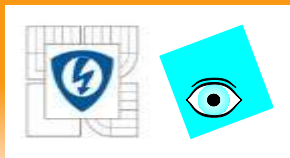




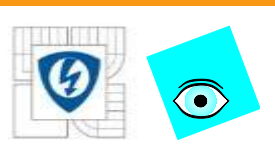
- (a) Pro symetrická rozdělení má kvantilová funkce sigmoidální tvar
- (b) Pro rozdělení sešikmená k vyšším hodnotám je konvexně rostoucí
- (c) Pro rozdělení sešikmená k nižším hodnotám je konkávně rostoucí
- (d) Zakreslují se tři obdélníky F, E a D



- (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: 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 $\tilde{x}_{0.5} \pm 1.57 \cdot R_F / \sqrt{n}$.

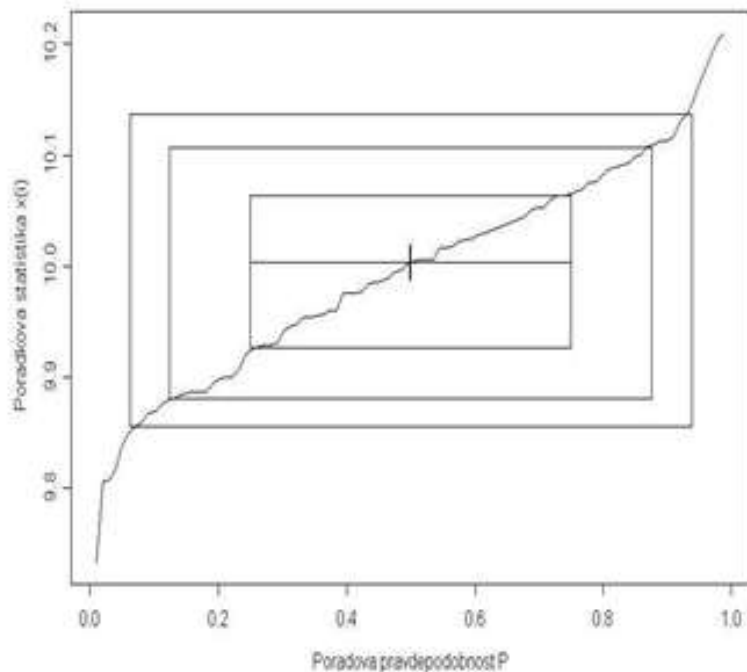


1. Symetrické unimodální rozdělení výběru: obdélníky symetricky uvnitř sebe.
2. Nesymetrická rozdělení: pro sešikmené rozdělení rozdílné vzdálenosti mezi dolními a horními hranami obdélníků.
3. Odlehlá pozorování: náhlý vzrůst na křivce, kdy směrnice roste nade všechny meze. Body vně sedecilového obdélníku.

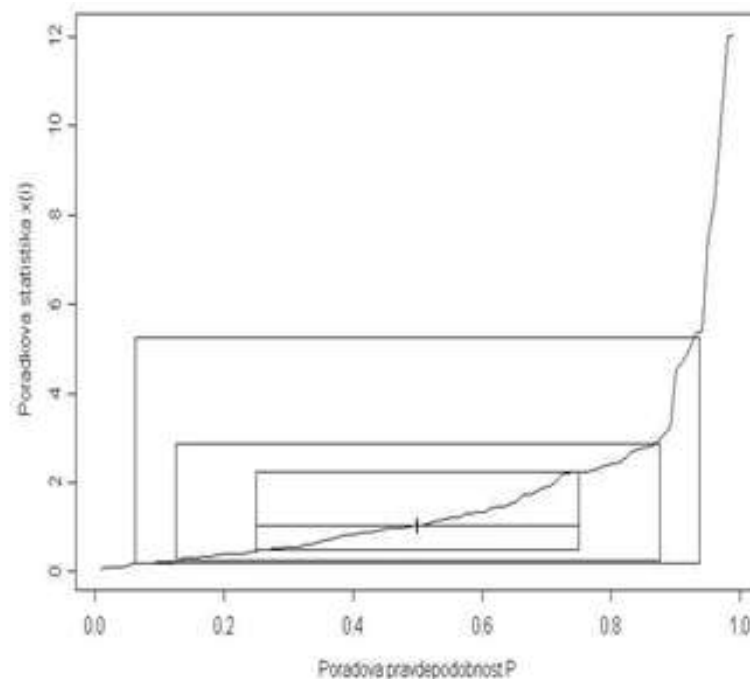


Graf rozptýlení s kvantily

G10



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

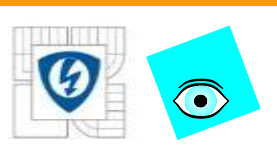


Graf rozptýlení s kvantily logaritmicko-normálního rozdělení *log* $LN(5, 2)$

31.1.2011

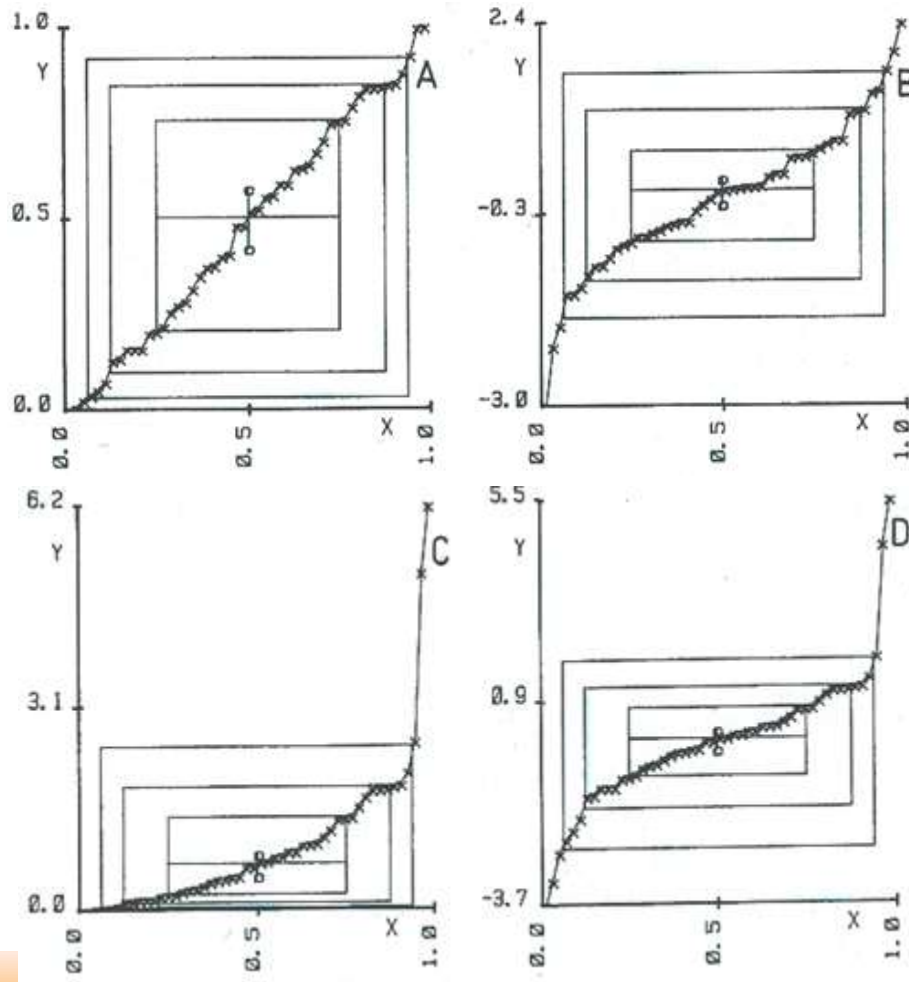
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





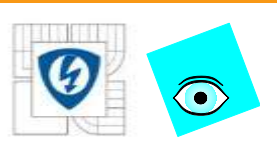
Graf rozptýlení s kvantily pro výběry z G10

(A) Rovnoměrného, (B) Normálního, (C) Exponenciálního a (D) Laplaceova



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

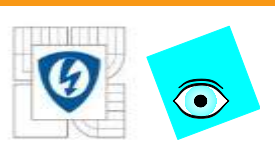


Graf rozptýlení s kvantily

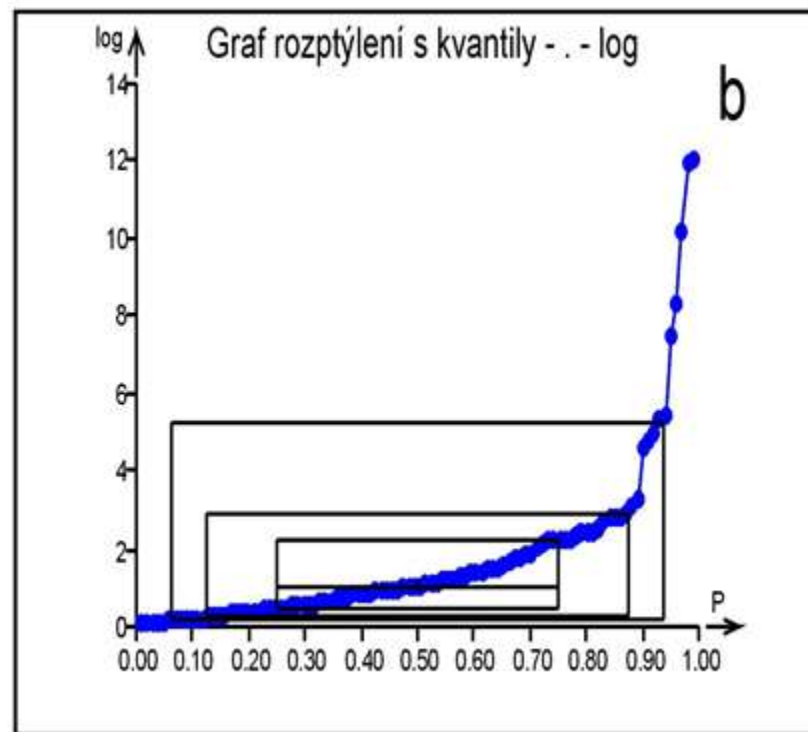
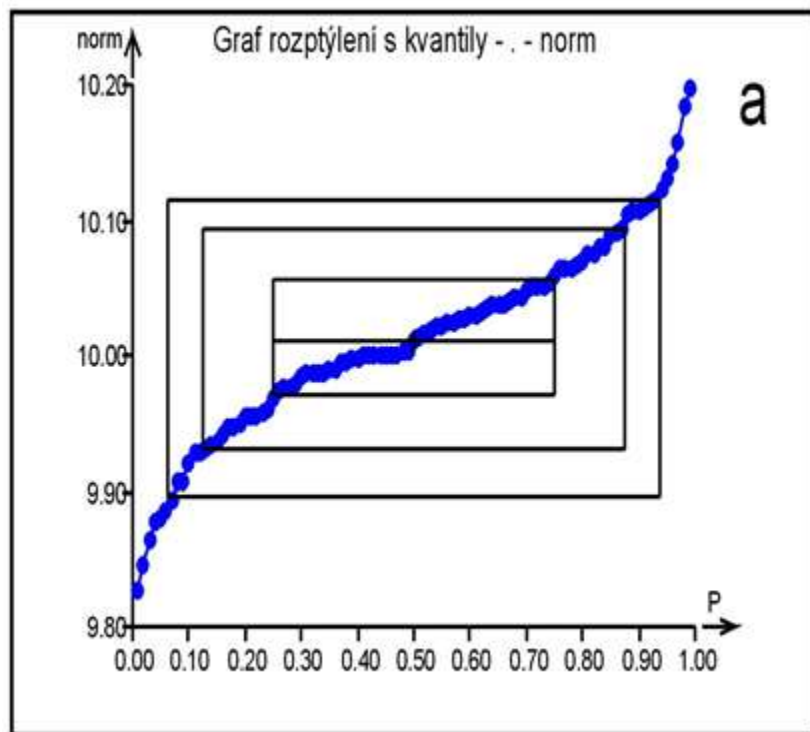
G10

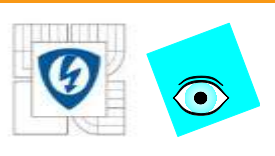
Osa x : P_i , Osa y : $x_{(i)}$

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.



(a) norm, symetrického (Gaussova, normálního), a (b) log, asymetrického (logaritmicko-normálního) rozdělení, ADSTAT.





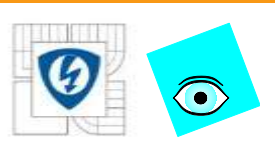
Identifikace rozdělení výběru

Posouzení zvláštností rozdělení výběru:

- (1) Jádrový odhad hustoty pravděpodobnosti
- (2) Histogram (resp. polygon, rootogram, atd.)

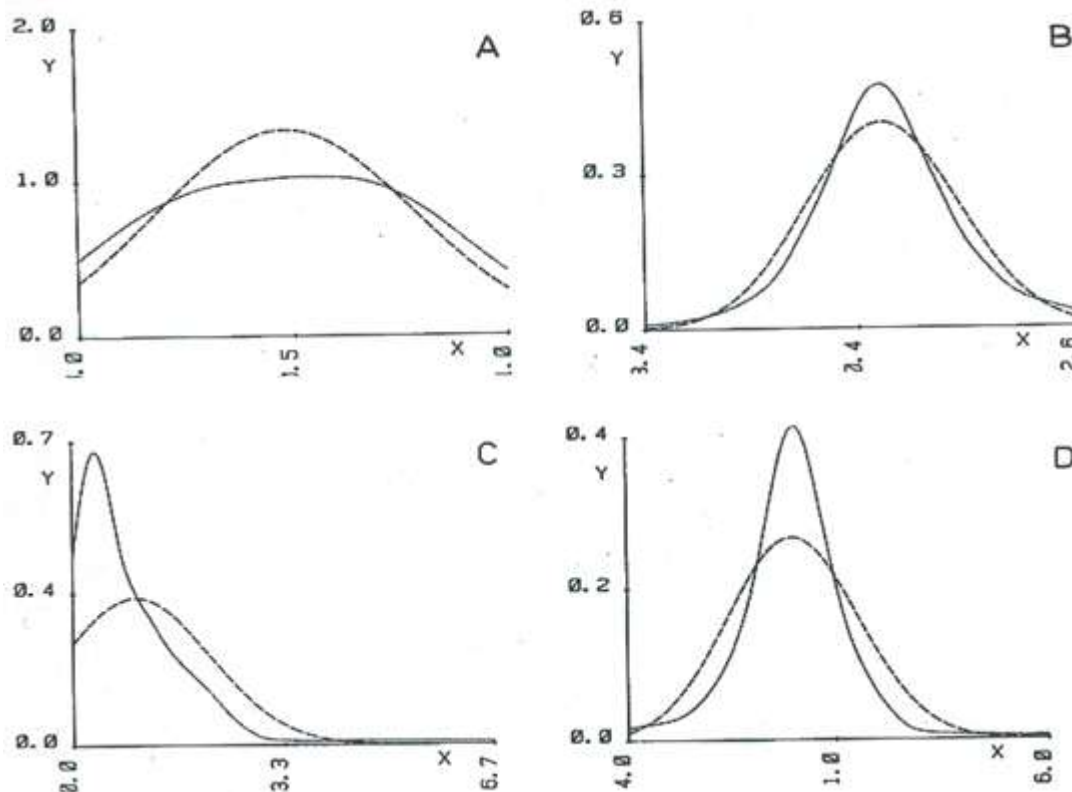
Porovnání rozdělení výběru s teoretickým rozdělením

- (1) Kvantil-kvantilový (Q-Q) graf
- (2) Rankitový graf
- (3) Podmíněný rankitový graf
- (4) Pravděpodobnostní graf
- (5) Kruhový graf



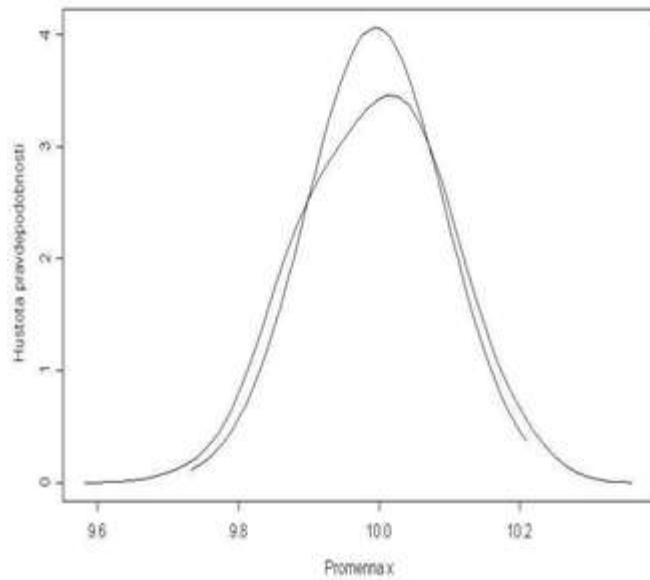
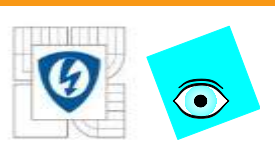
Jádrový odhad hustoty pravděpodobnosti G11

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

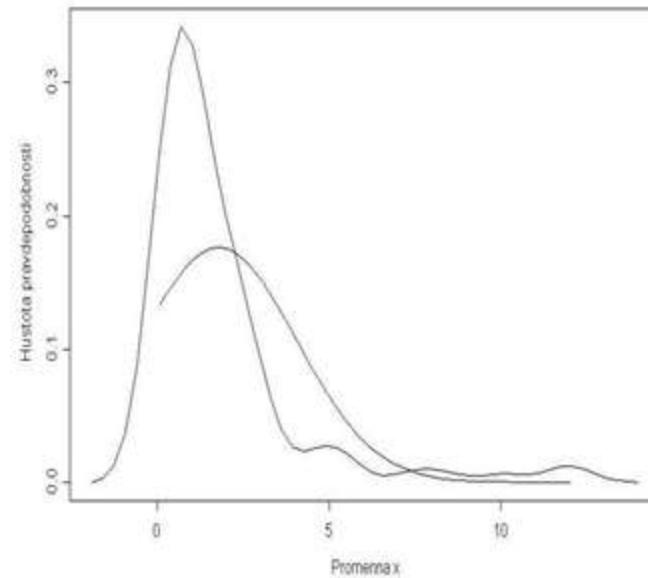


Čárkovaně je znázorněna hustota Gaussova rozdělení s parametry $\bar{x}$ a s^2 a plnou čarou járový odhad hustoty pravděpodobnosti empirického rozdělení výběru.

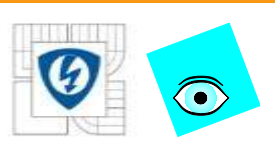
Jádrové odhady hustoty pravděpodobnosti pro výběry z rozdělení
(A) Rovnoměrného, (B) Normálního, (C) Exponenciálního a (D) Laplaceova



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



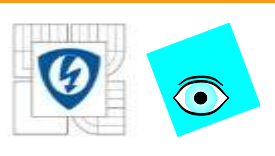
Graf hustoty pravděpodobnosti logaritmicko-normálního rozdělení **log LN** $(5, 2)$



1. Zkonstruuje se předběžný odhad $\hat{f}_0(x)$ se šířkou pásu
$$h_0 = 0.75 \left(\frac{n}{100} \right)^{-0.2} \min_{(i)} [x_{(i+\text{int}(\frac{n}{2}))} - x_{(i)}]$$
2. Sestrojí se vlastní odhad hustoty pravděpodobnosti $\hat{f}_K(x)$ s využitím jádrové funkce ale s nekonzistentní 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 šířku pásu platí
$$h_i = h_0 \left[\frac{\hat{f}_0(x_i)}{\max_{(i)} \hat{f}_0(x_i)} \right]^{-\alpha}.$$

Parametr $\alpha \in (0,1)$ ovlivňuje hladkost odhadu $\hat{f}(x)$

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

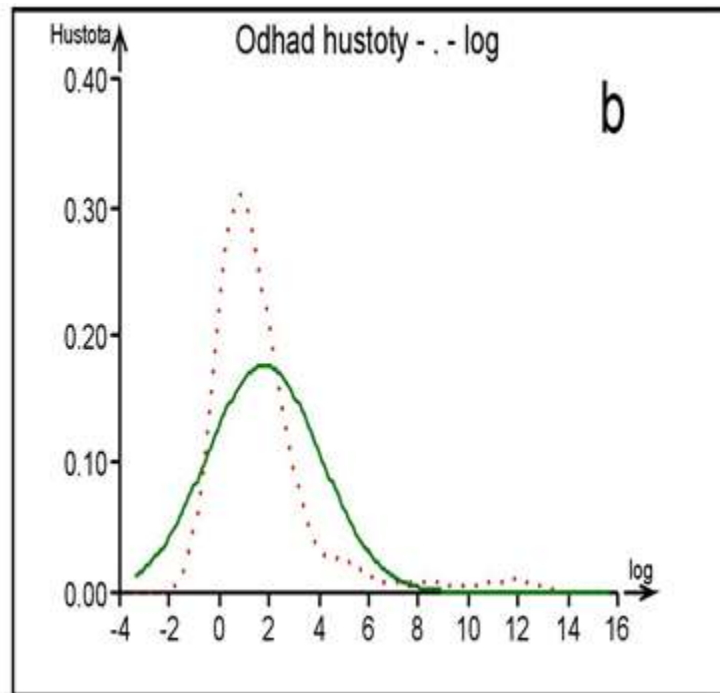
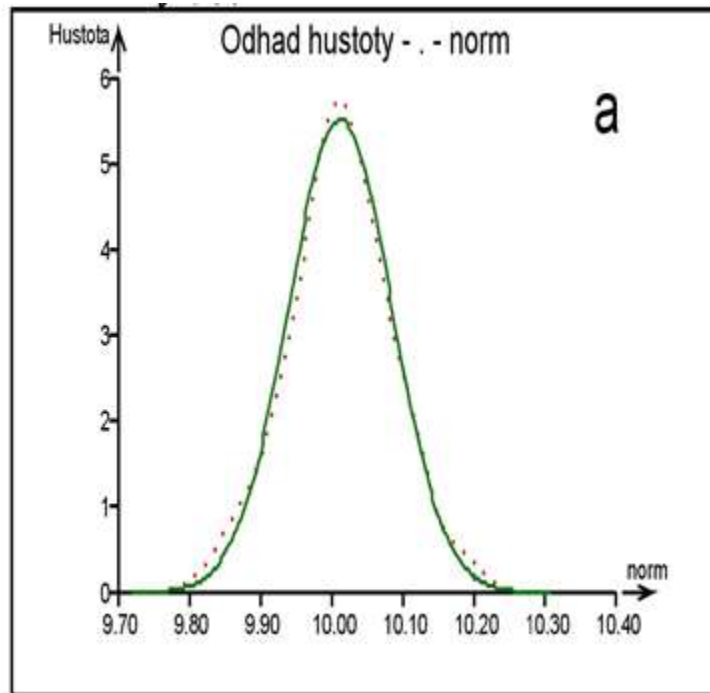


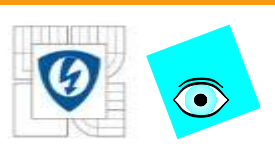
Jádrový odhad hustoty pravd.

G11

Osa x : proměnná x , Osa y : hustota pravděpodobnosti $\hat{f}(x)$

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



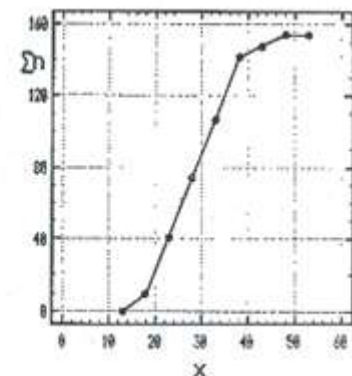
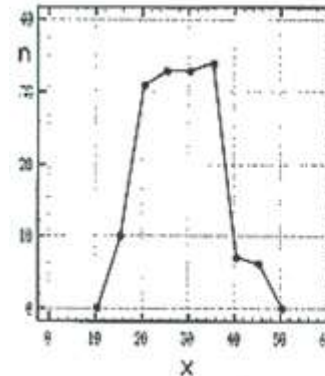
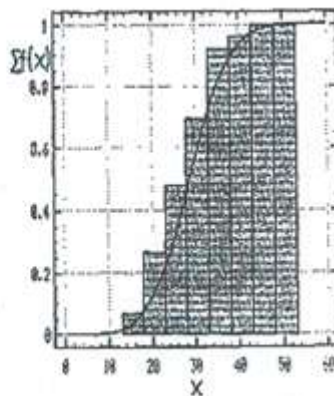
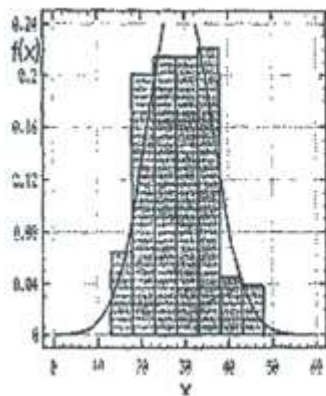


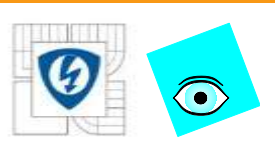
Histogram

G12

Osa x : proměnná x , Osa y : hustota pravděpodobnosti $\hat{f}(x)$

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





Délka třídního intervalu :

$$\Delta x_j = x_{j+1}^+ - x_j^+$$

Volba počtu tříd L :

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

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

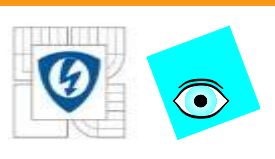
Optimální délka třídních intervalů: $\Delta x_{\text{opt}} = 3.49 \frac{s}{n^{1/3}}$

kde s je směrodatná odchylka.

Robustní odhad délky třídních intervalů:

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

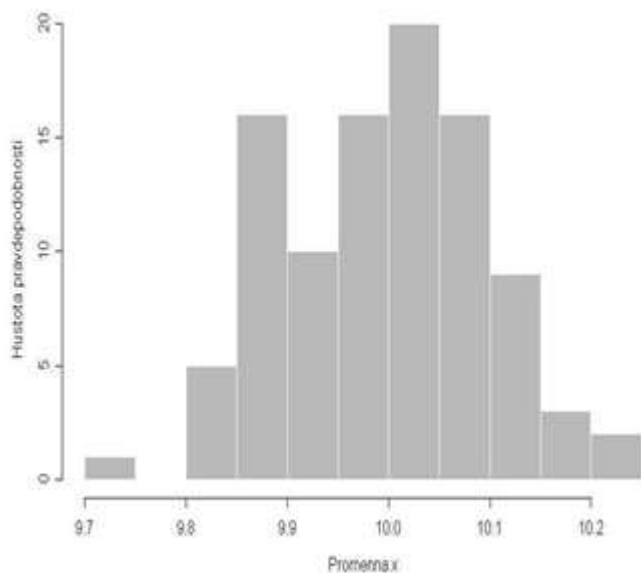
kde F_H a F_D jsou výběrové kvartily.



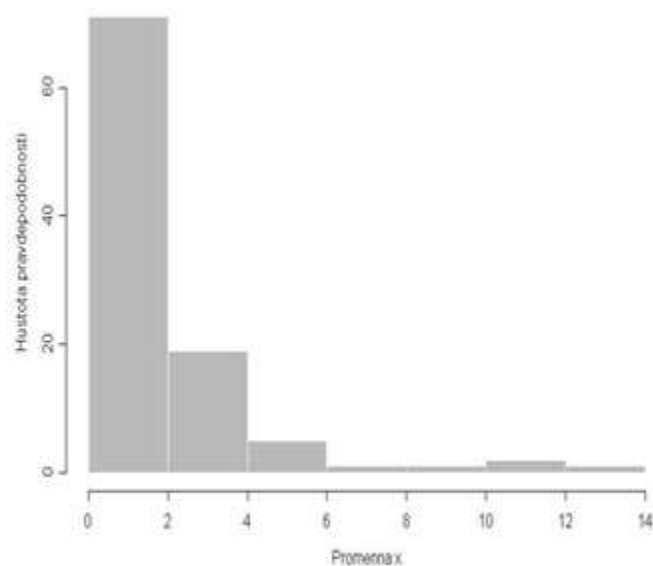
Histogram

G12

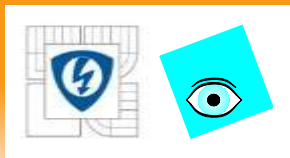
Histogram



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



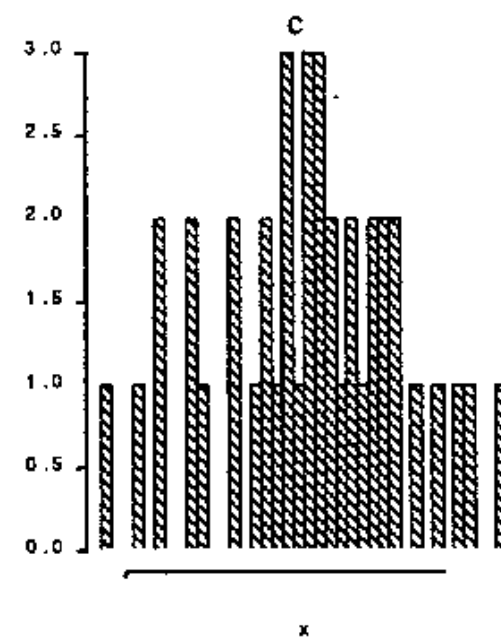
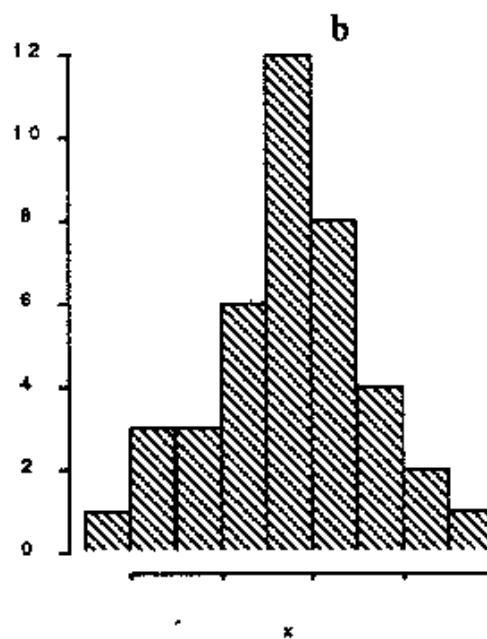
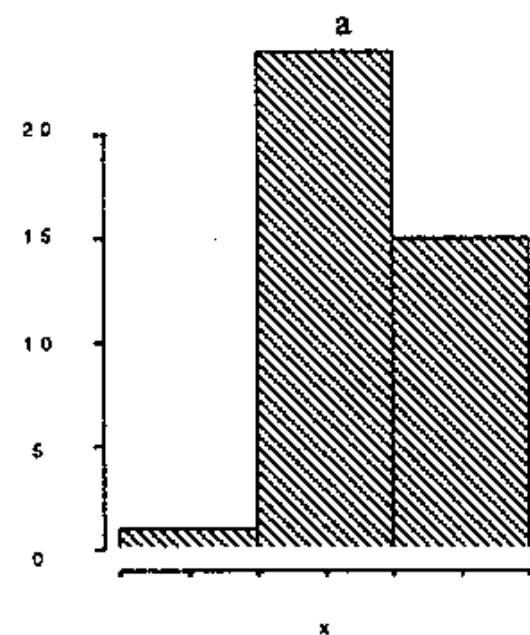
Histogram logaritmicko-normálního rozdělení *log* $LN(5, 2)$

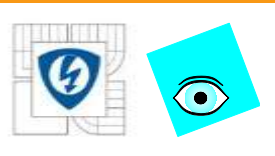


Důležitá je správná volba šířky třídy:

$$L = \text{int}[2.46(n - 1)^{0.4}]$$

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



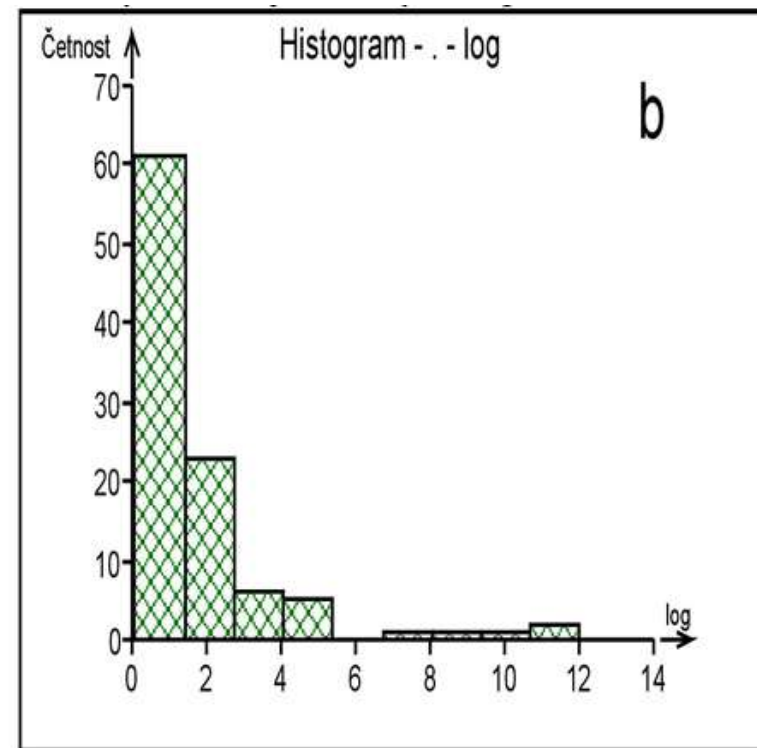
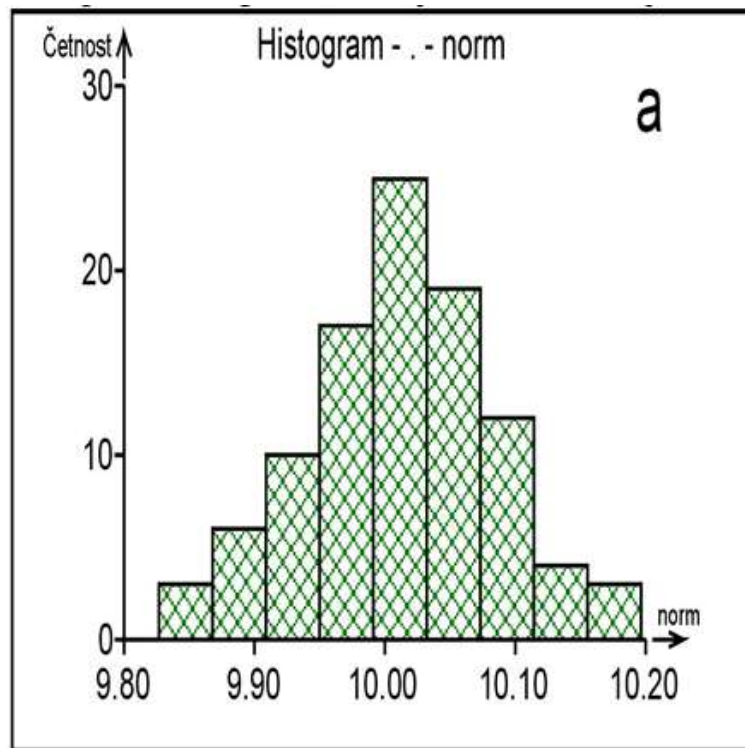


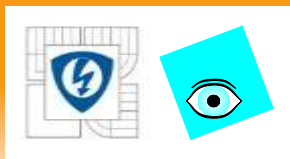
Histogram

G12

Osa x : proměnná x , Osa y : =úměrná hustotě pravděpodobnosti

Jde o obrys sloupcového grafu, kde jsou na ose x jednotlivé třídy, definující šířky sloupců.





Kvantil-kvantilový graf (graf Q-Q)

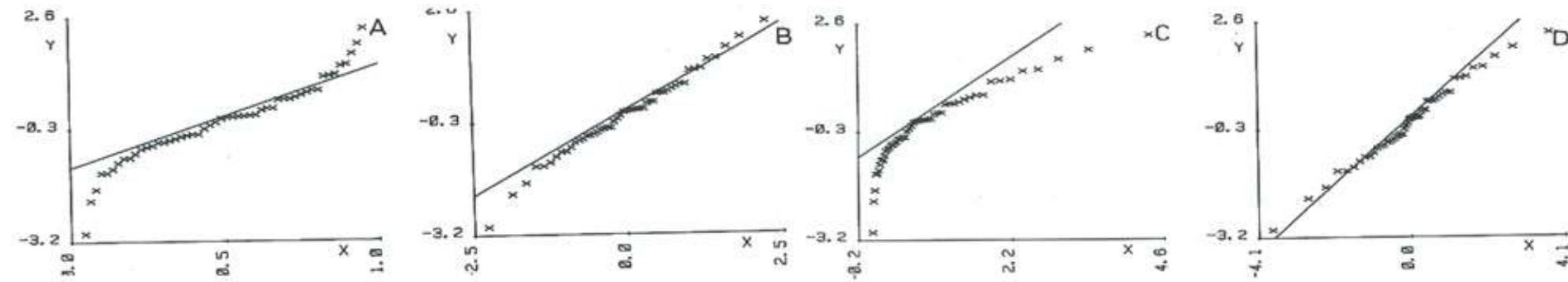
G13

Osa x : $Q_s(P_i)$, Osa y : $x_{(i)}$

Grafy Q-Q pro porovnání rozdělení výběru rozdělení s teoretickým (A) Rovnoměrným, (B) Normálním, (C) Exponenciálním a (D) Laplaceovým rozdělením. Čárou je vyznačen teoretický průběh.

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

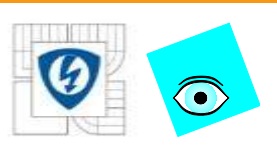
kde P_i je pořadová posloupnost. Při shodě rozdělení s teoretickým, je $x_{(i)}$ na $Q_T(P_i)$ lineární funkcí.



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





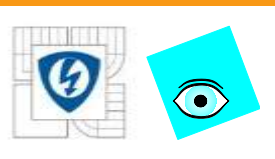
Obvykle lze normovat použitím proměnné

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

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

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

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

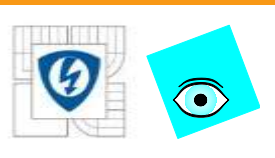


Kvantil-kvantilový graf (graf Q-Q)

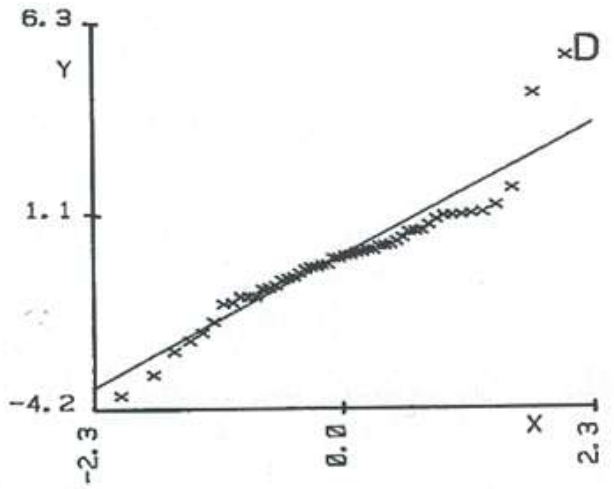
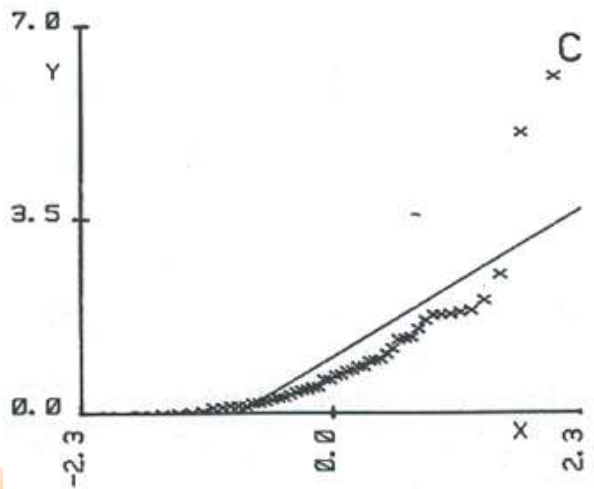
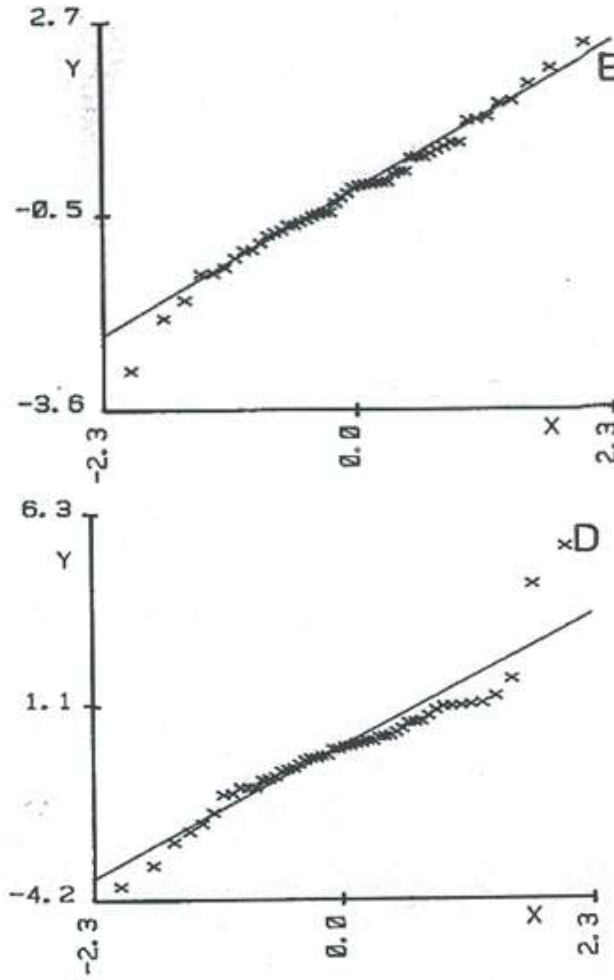
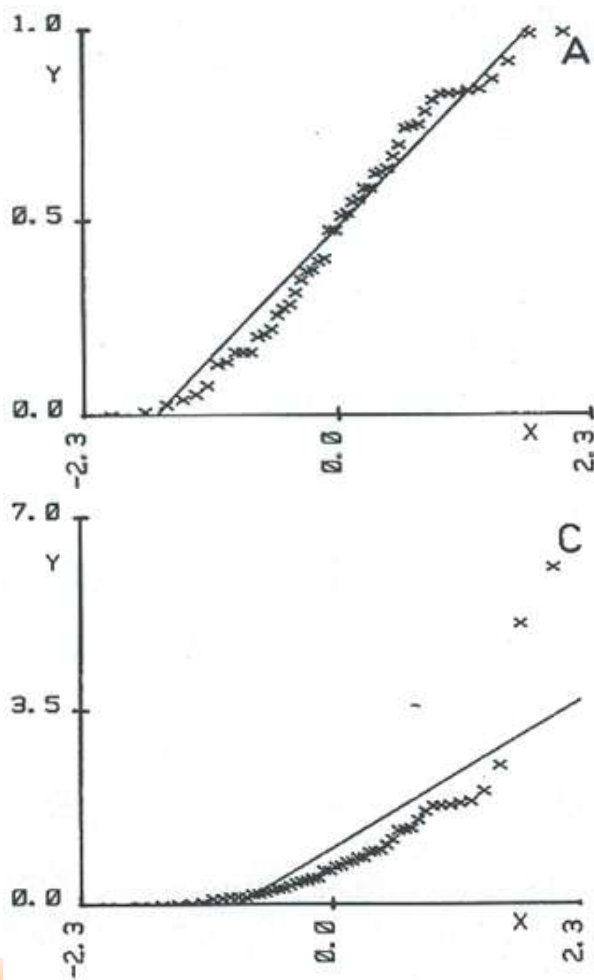
G13

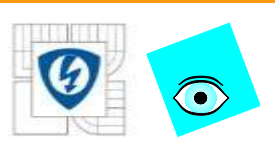
Standardizované hustoty a distribuční funkce vybraných rozdělení a odpovídající souřadnice grafů Q-Q

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	$\Phi[\ln(s)]$	$\frac{\exp[-0.5 \ln(s)^2]}{\sqrt{2 \pi}}$	$x_{(i)}$	$\exp[\Phi^{-1}(P_i)]$



(A) Rovnoměrného, (B) Normálního, (C) Exponenciálního a (D) Laplaceova



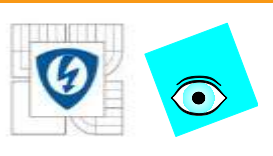


Kvantilově-kvantilový graf (grafQ-Q)

G13

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.

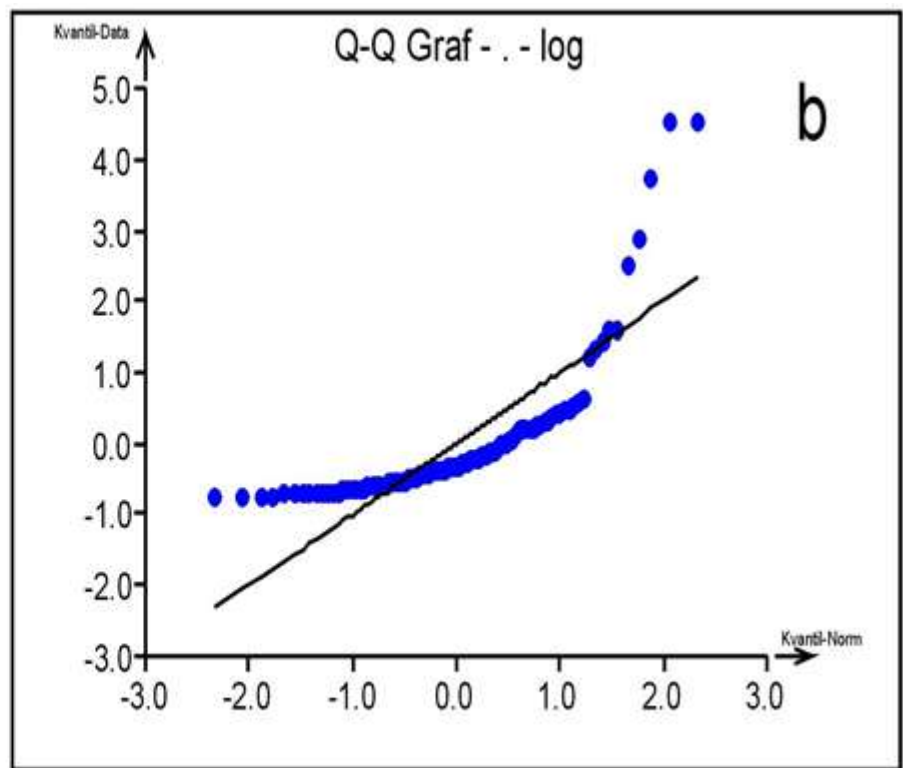
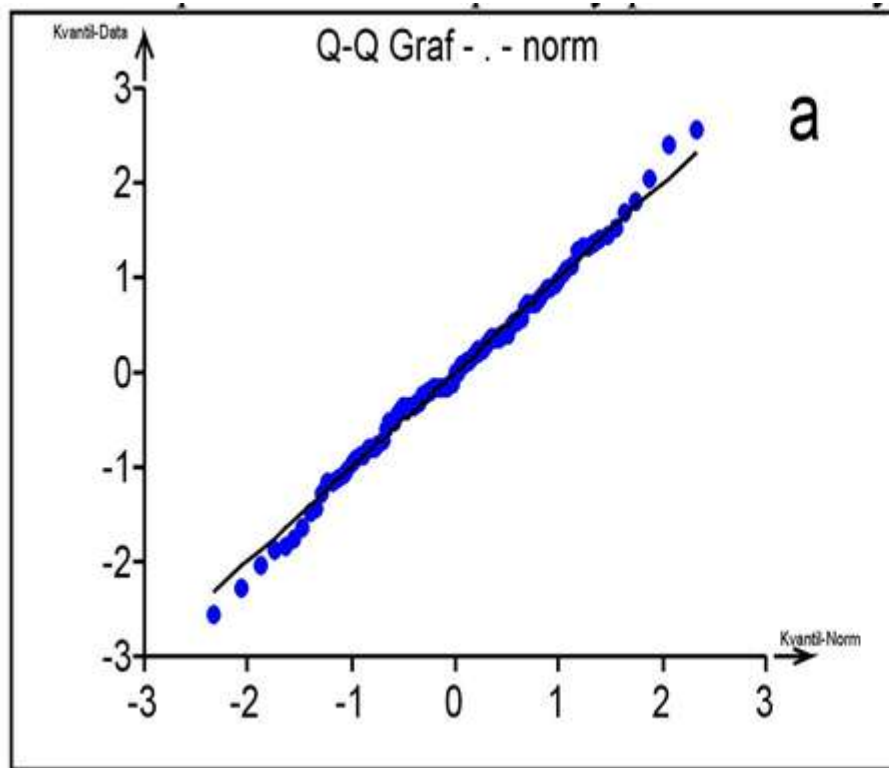


Kvantilově-kvantilový graf (grafQ-Q)

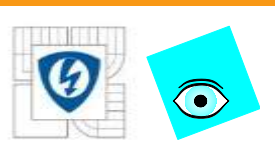
G13

Osa x : $Q_T(P_i)$, Osa y : $x_{(i)}$

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.

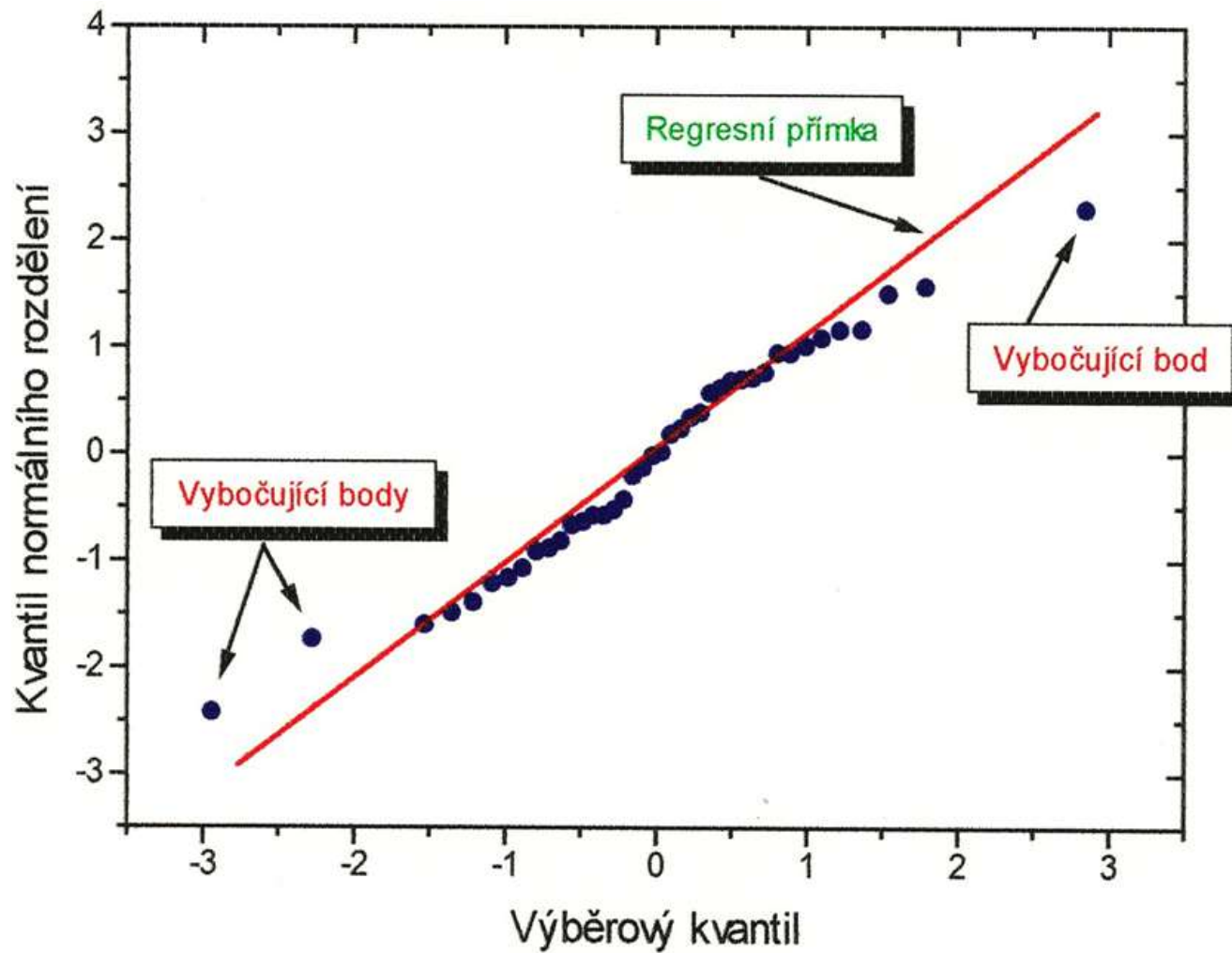


31.1.2011



Q-Q graf

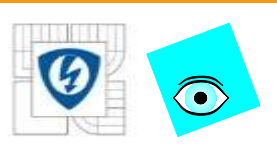
G14



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



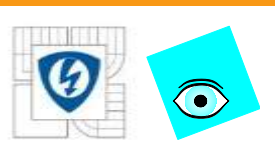


K porovnání rozdělení výběru s rozdělením normálním se Q-Q graf nazývá **rankitový graf**.

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

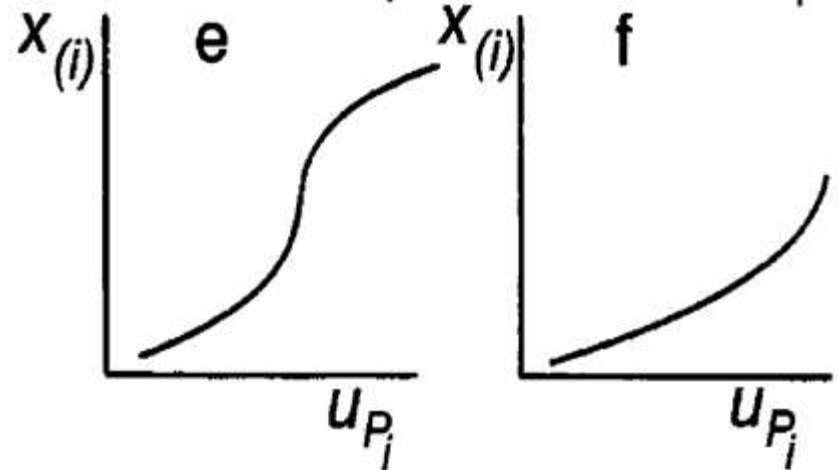
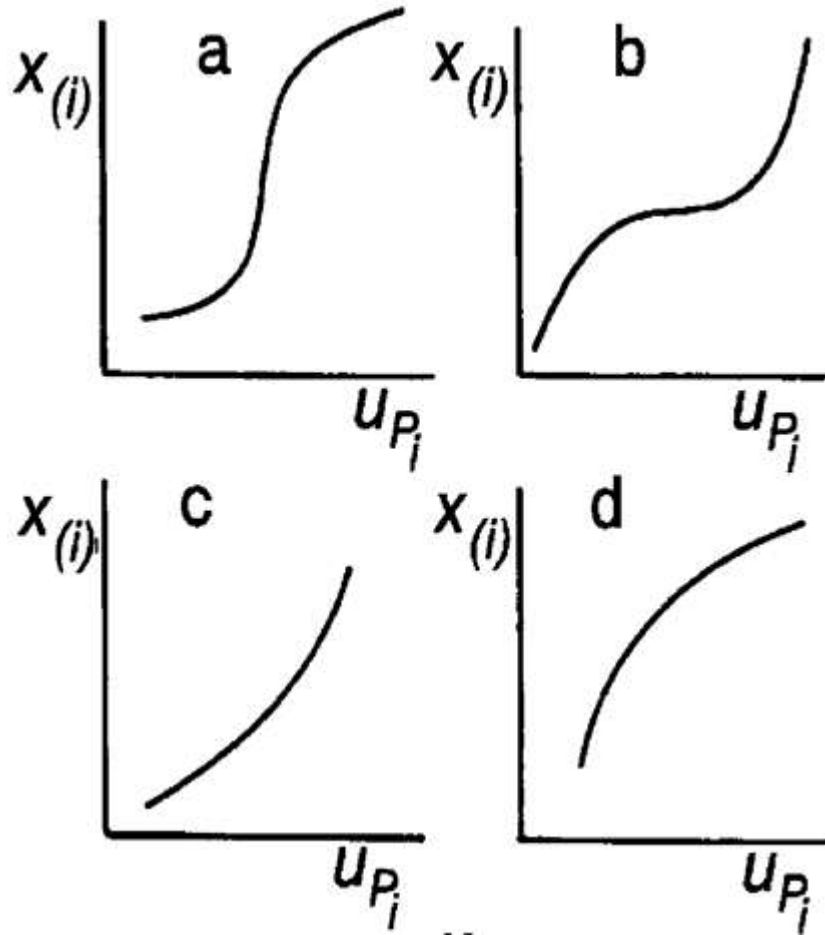
Konvexní, popř. konkávní průběh Q-Q grafu indikuje zešikmené rozdělení výběru, zatímco esovitý průběh ukazuje na rozdíl v délce konců ve srovnání s normálním rozdělením.

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



Rankitový graf

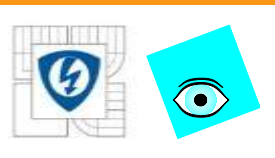
G14



31.1.2011

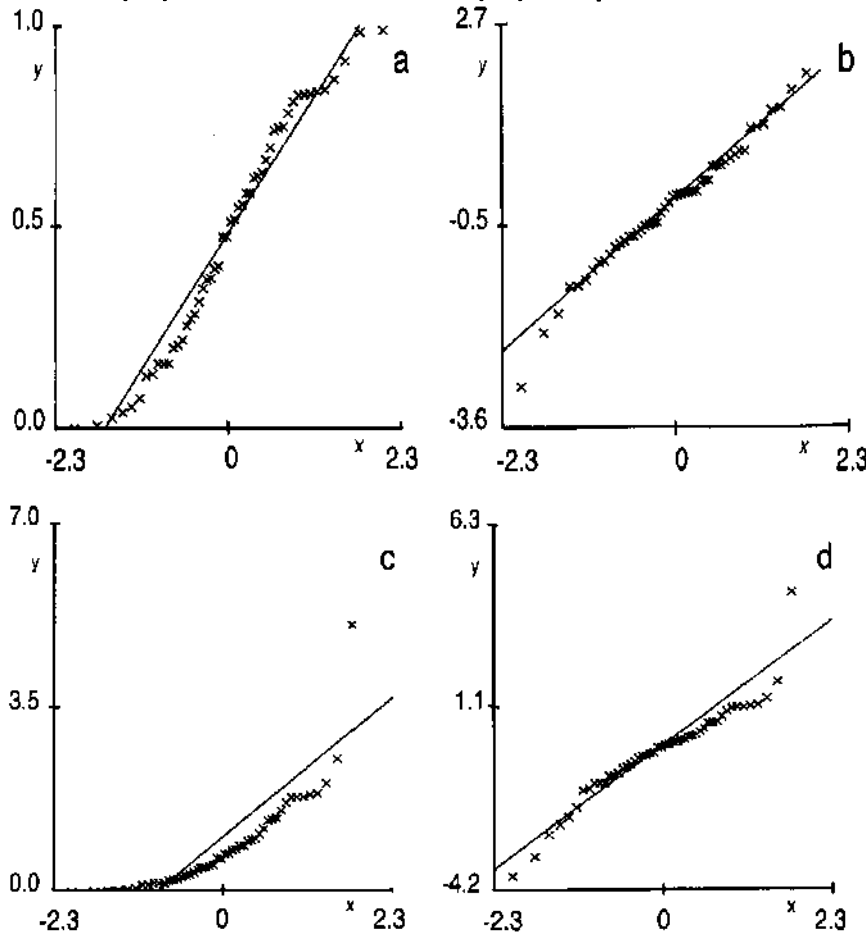
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



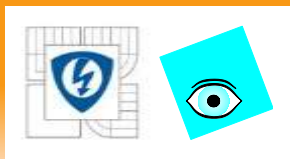


Rankitové grafy pro výběry z rozdělení G14

(A) Rovnoměrného, (B) Normálního, (C) Exponenciálního a (D) Laplaceova



Plnou čarou je vyznačen teoretický průběh.



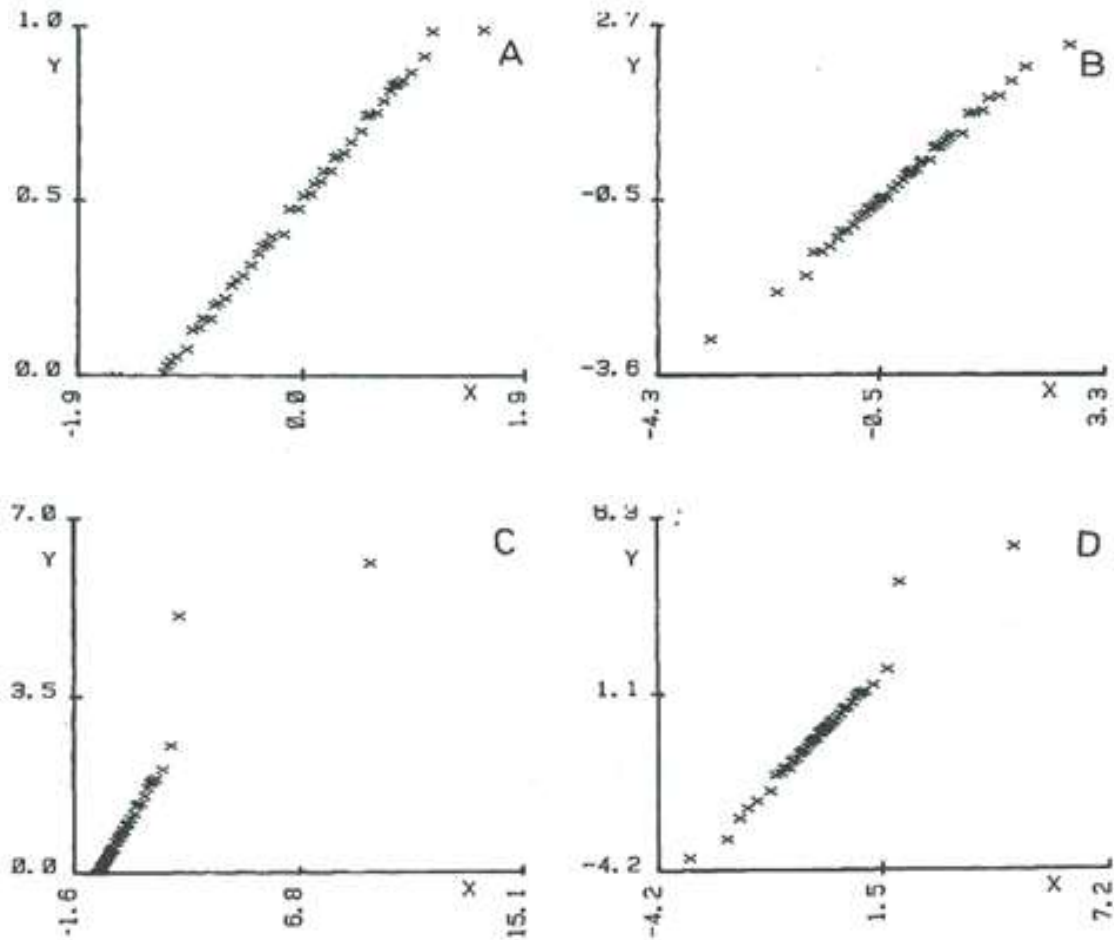
Podmíněný rankitový graf

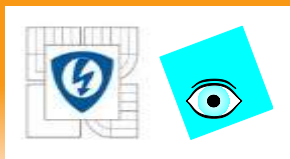
G15

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

Pro výběry z rozdělení
(A) rovnoměrného,
(B) normálního,
(C) exponenciálního a
(D) Laplaceova

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





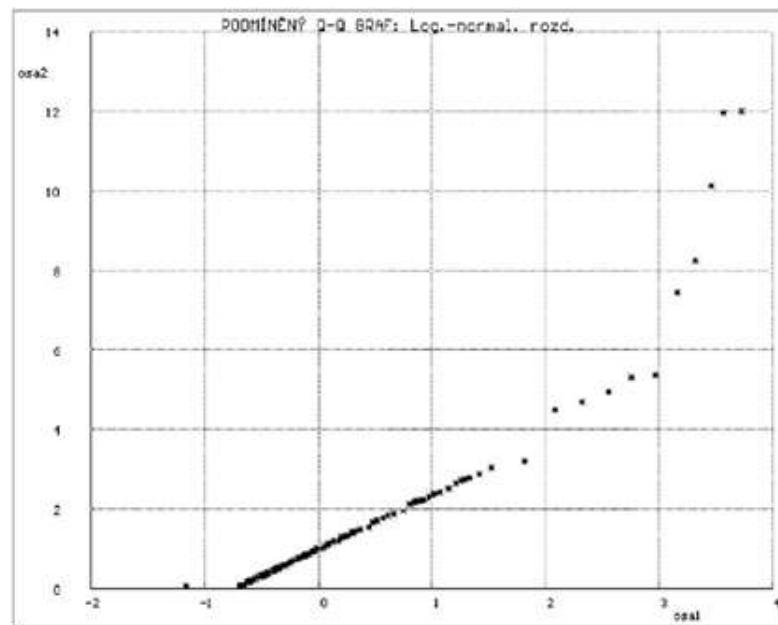
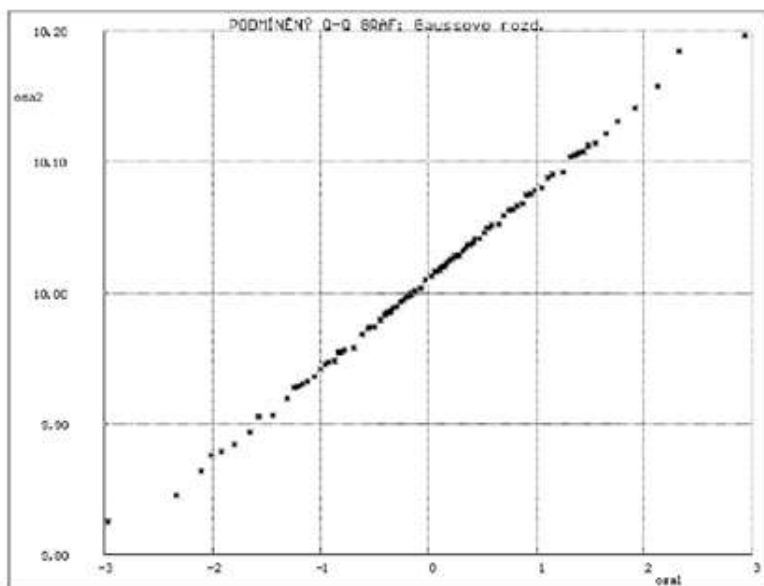
Podmíněný rankitový graf

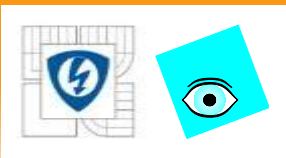
G15

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

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



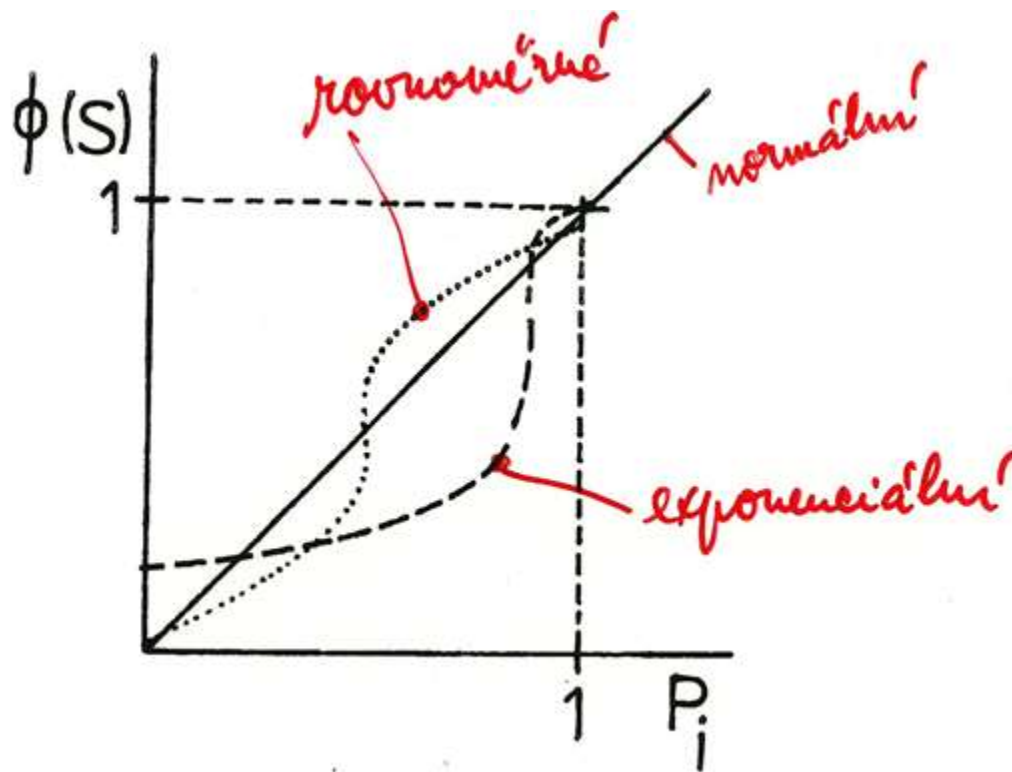


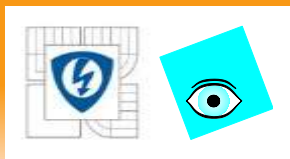
Pravděpodobnostní graf (P-P graf)

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.





Osa x : P_i , Osa y : $F_T(S_{(i)})$

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.

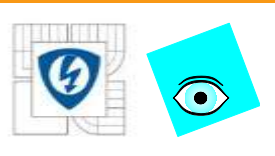


Diagnóza:

- (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:

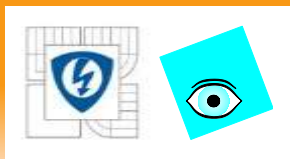
1. Osa x : $P_i = i/(n + 1)$, osa y : $\Phi[(x_{(i)} - \bar{x})/s]$,
2. Výběr k hodnot (cca $k = 200$) a výpočet $P_i = \frac{i}{k+1}$



3. Generace standardizovaných kvantilů $Q_{si} = G^{-1}(P_i)$ a Q, R pro rozdělení $F_T(.)$. Pro normální rozdělení

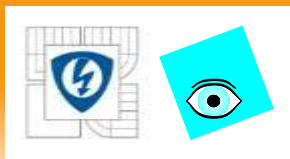
$$\hat{Q} = \frac{1}{k} \sum_{i=1}^k Q_{si} \quad \text{a} \quad \hat{R} = \sqrt{\frac{1}{k} \sum_{i=1}^k (Q_{si} - \hat{Q})^2}$$

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



Osa x : P_i , Osa y : $F_T(S_{(i)})$

Slouží k porovnání distribuční funkce výběru, vyjádřené přes požadovanou 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.

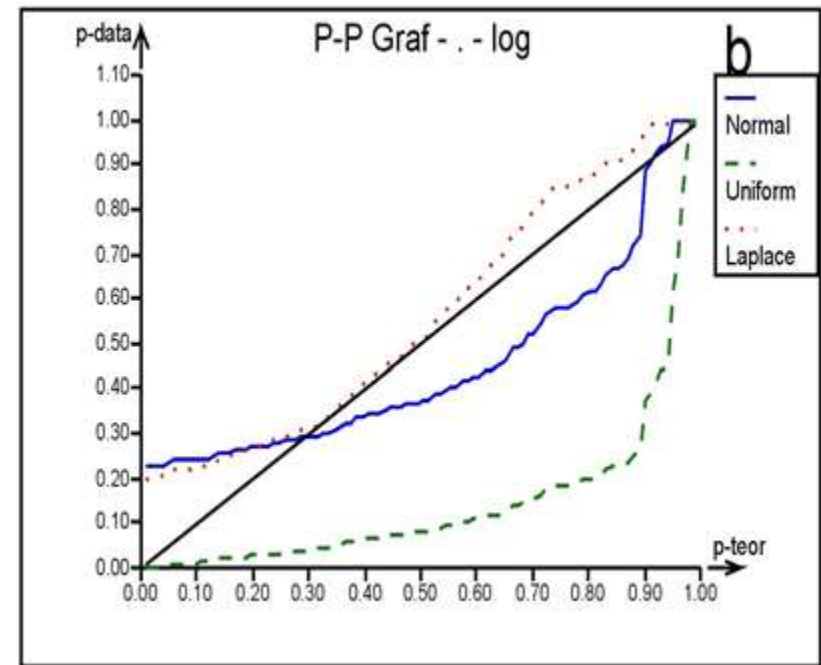
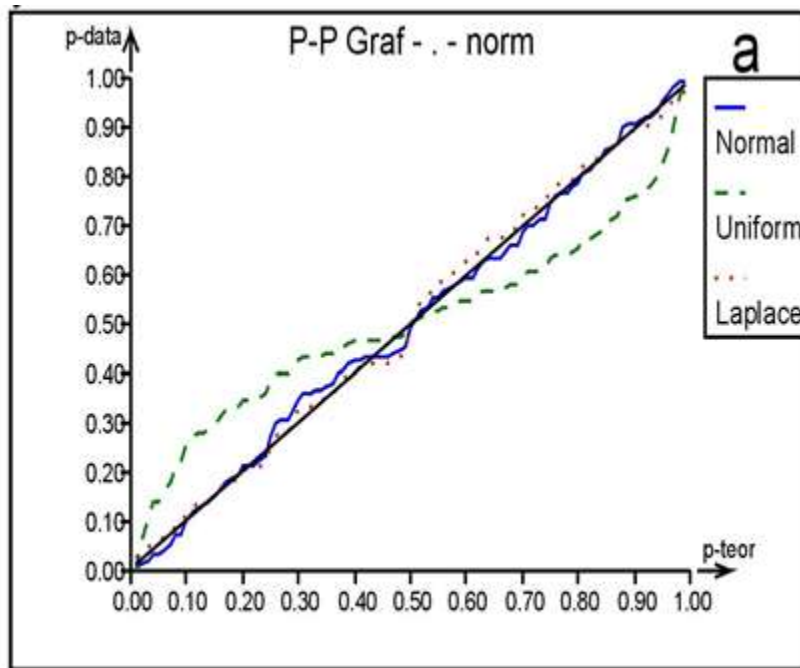


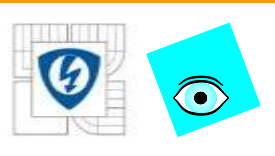
Pravděpodobnostní graf (P-P graf)

G16

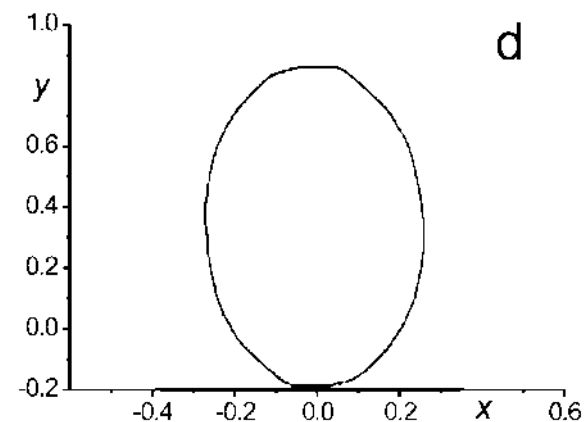
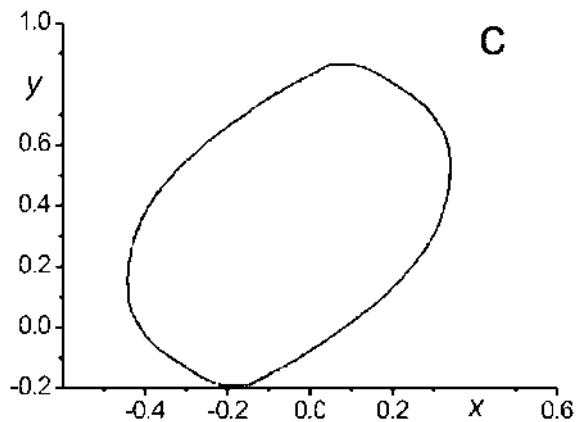
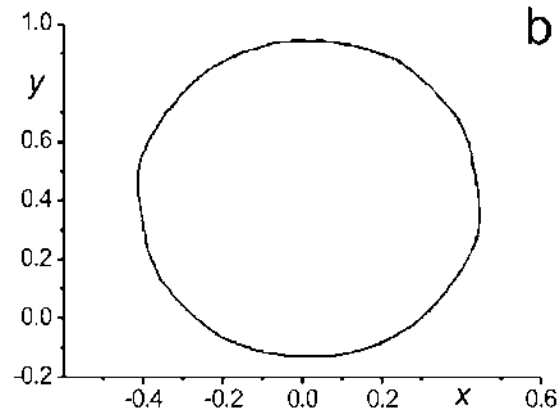
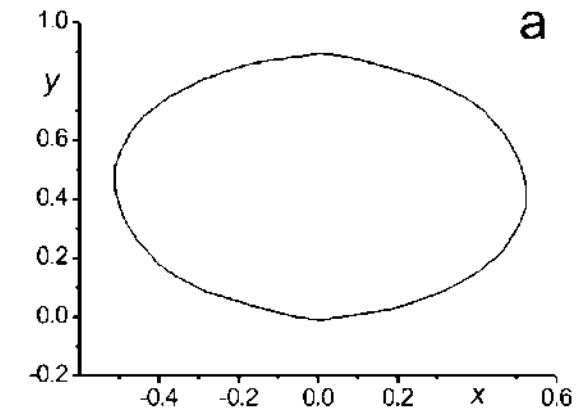
Osa x : P_i , Osa y : $F_T(S_{(i)})$

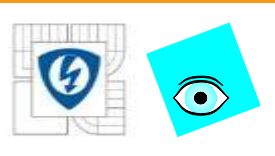
Pravděpodobnostní graf (P-P graf) pro výběry: (a) norm, symetrického (Gaussova, normálního), a (b) log, asymetrického (logaritmicko-normálního) rozdělení, ADSTAT.



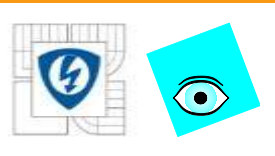


Kruhový graf pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova





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



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

