EXDAB - programový systém pro průzkumovou analýzu dat

Jiří Militký, VÚTZ Dvůr Králové n.L.

1. Úvod

V literatuře existují dva krajní přístupy ke zpracování dat. Při datově analytickém přístupu se na základě analýzy experimentálních údajů hledají nové informace. Centrální roli zde hrají data, o nichž se předpokládá, že obsahují nové poznatky.

Přístup vědecko-poznávací vychází z modelu sledovaného problému vytvořeného z předchozích informací. Data zde slouží jako prostředek pro ověření modelu resp. jeho rozšíření.

V praxi se běžně oba přístupy kombinují a hledají se informace skryté v datech i vhodné pravděpodobnostní modely.

Zde se zaměříme na nejfrekventovanější úlohu zpracování jednorozměrných výběrů. Při datově analytickém přístupu se postup rozpadá do dvou fází:

- exploratorní analýza dat (EDA)
- konfirmatorní analýza dat (CDA)

Postup CDA slouží pro "zodpovězení" otázek stanovených na základě EDA. Umožňuje vytváření pravděpodobnostních modelů a jejich testaci. Zde se dále budeme zabývat pouze první fází, tj. EDA.

Pod pojmem EDA (označovaný také jako průzkumová analýza dat) se skrývají techniky, umožňující analýzu struktury a zvláštností dat.
J.W. Tukey /1/, duchovní otec metod průzkumové analýzy dat, zahrnuje do této oblasti:

- všechny prostředky pro ulehčení popisu a sumarizace dat
- všechny prostředky pro hlubší pochopení statistických souvislostí v datech

Jednotlivé techniky využívají pouze základních pravděpodobnostních předpokladů (spojitost a diferencovatelnost frekvenční funkce, atp.). Jejich specifikou je používání robustních (rezistentních) metod a vlastní "nestatistické" názvosloví. Jsou vhodné zejména pro případy, kdy se očekává nalezení nových struktur.

Pro případ rutinní analýzy dat se místo postupů EDA doporučuje využití metod IDA (Initial Data Analysis). Pod pojmem IDA se skrývají všechny postupy umožňující stanovení cílů analýzy, popis dat a předběžnou formulaci pravděpodobnostního modelu. Kromě technik exploratorního charakteru se využívá i metod pro testaci nezávislosti, normality, homogenity výběru, atp., které již vycházejí z pravděpodobnostních modelů.

Programový soubor EXDAB (EXploratory DAta Analysis in Basic), popsaný v tomto příspěvku, obsahuje vybrané techniky IDA a EDA se zaměřením na grafické procedury. Umožňuje řešení těchto základních problémů:

a) vyjádření dat tak, aby vynikly jejich typické statistické zvláštnosti (program "EX-GR")
b) orientační posouzení shody empirického rozdělení výběru se zvoleným teoretickým rozdělením (program "GR-POR")
c) transformace dat tak, aby došlo ke zlepšení statistických vlastností výběru (symetrie, délka konců, stabilizace rozptylu a ideálně normality) - program "TRANSF".

Činnost souboru EXDAB je demonstrována na příkladu z práce /9/. Jde o 30 údajů o velikosti dešťových srážek v průběhu jednoho měsíce. Základní statistické

údaje jsou uvedeny v tab. 1.

Tab. 1 Základní statistické údaje o výběru z /9/.

průměr $\bar{x}$	1.67	rozptyl s^2	1.C01	šikmost $\hat{g}_1$	1.087
medián $\tilde{x}$	1.47	módus $\hat{x}$	1.125	špičatost $\hat{g}_2$	4.207

Na tomto příkladě lze velmi ilustrativně sledovat, jak se postupně "objevují" důvody pro zpochybnění předpokladu normality.

2. Grafy pro znázornění zvláštností v datech

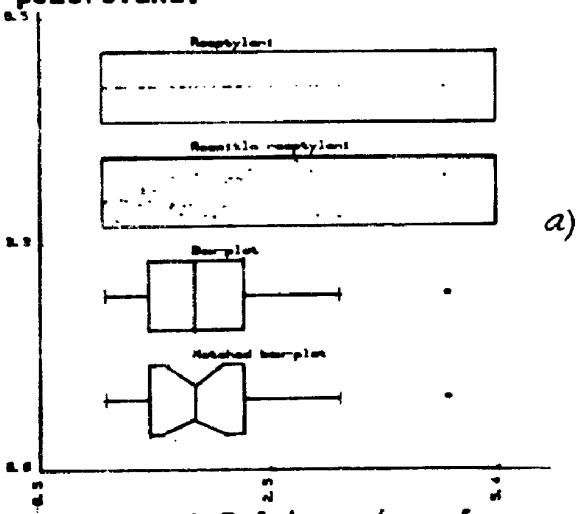
Do této skupiny patří řada velmi jednoduchých postupů pro částečnou sumarizaci dat resp. jejich vizualizaci. V programu "EX-GR" je použito těchto grafů:

1. Grafy rozptýlení a rozmítlé rozptýlení
2. Krabicové grafy a grafy rozptýlení s kvantily
3. Kvantilové grafy
4. Grafy symetrie
5. Grafy empirické frekvenční funkce a histogramy

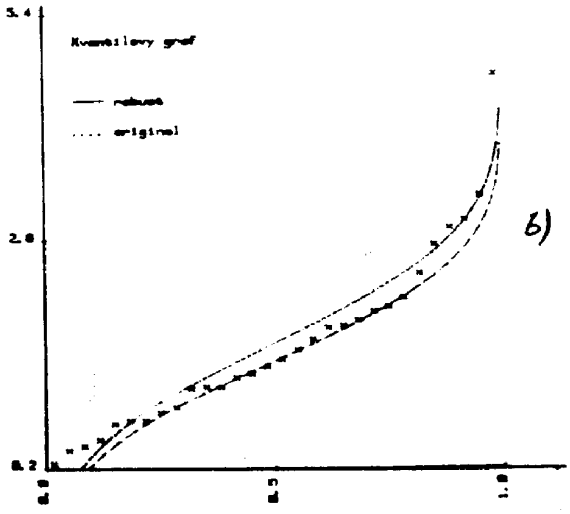
Protože se předpokládá využití těchto technik zejména pro malé a střední výběry, využívá se pořádkových statistik $x_{(i)}$ i=1,...n (platí $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$). Základní vlastnosti pořádkových statistik jsou uvedeny v kap. 3.

2.1 Grafy rozptýlení a rozmítlé rozptýlení

To je nejjednodušší způsob vizualizace dat, spočívající v jejich vynesení na reálné ose. U klasického grafu rozptýlení se vynáší při stejné hodnotě y= = konst. body $\{x_i, y\}$ i=1,...n. Jde vlastně o projekci kvantilového grafu (viz 2.3). Pro zamezení splývání dat se používá "rozmítlého" rozptýlení, kdy se místo y=konst. provádí vertikální rozptýlení jednotlivých bodů (v intervalu $\langle 0,1 \rangle$) pomocí generátoru pseudonáhodných čísel. Detaily o interpretaci těchto grafů lze nalézt v práci /10/. Na obr. 1a je pro uvedený příklad znázorněn graf rozptýlení spolu s rozmítlou verzí. Je patrné sešikmení dat a přítomnost odlehlého pozorování.



a)



b)

Obr. 1 a) Průzkumové grafy
b) Kvantilový graf.

2.2 Krabicové grafy a grafy rozptýlení s kvantily

Krabicové grafy (Box-plot) patří mezi základní techniky průzkumové analýzy /1/. Umožňují sumarizaci většiny dat a přitom indikaci polohy, rozptýlení, délky konců a příp. i vybočujících pozorování. Je pro ně zavedena speciální terminologie /1/.

V nejjednodušší variantě je krabicový graf obdélník o šířce $\Delta_I = x_{0,75} - x_{0,25}$

(odpovídá interkvartilové vzdálenosti). V místě mediánu je vertikální čára. Od obou konců tohoto obdélníku pokračují vertikální přímky, které jsou ukončeny v místech odpovídajících přilehlým pozorováním (tj. pozorováním, ležícím uvnitř intervalu $[\tilde{x}_{0.25} - 1.5\Delta_I, \tilde{x}_{0.75} + 1.5\Delta_I]$ nejblíže k jeho krajům). Pozorování, ležící mimo tyto hranice, se znázorňují kroužky (vzdálená pozorování). Pro zhodnocení konfidenčního intervalu mediánu je vhodný vrubový krabicový graf. Délka vrubu zde odpovídá 95%nímu intervalu spolehlivosti pro medián, který je definován vztahem $\tilde{x}_{0.5} \pm 1.57\,\Delta_I/\sqrt{n}$ (viz /10/). Detaily o interpretaci těchto typů grafů lze nalézt v /10/.

Na obr. 1a jsou znázorněny jak krabicový graf, tak vrubový graf krabicový. Je patrná symetrie dat až do kvartilů a sešikmení v oblasti konců směrem k vyšším hodnotám.

Zobecněním krabicových grafů jsou grafy rozptýlení s kvantily (GRK). Tyto grafy umožňují ve vhodné formě posoudit řadu zvláštností rozdělení dat. Protože mohou indikovat také typ rozdělení výběru, jsou diskutovány v kap. 3.1.

2.3 Kvantilové grafy

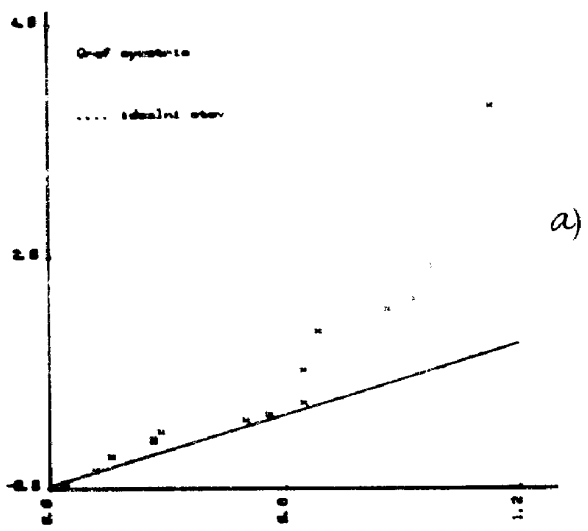
Jak plyne z poznámky v kap. 3, jsou pořádkové statistiky $x_{(i)}$ odhadem kvantilové funkce výběru $Q_E(P_i)$, $P_i = i/(n+1)$. Kvantilový graf se tedy jednoduše sestrojí vynesením hodnot $x_{(i)}$ na osu y a odpovídajících hodnot P_i na osu x. Z tohoto grafu lze usuzovat na symetrii, sešikmení, lokální koncentraci dat a orientačně i na normalitu.
Pro usnadnění interpretace se do grafu vynáší také kvantilové funkce normálního rozdělení x_p, pro které je $x_p = \hat{\mu} + \hat{\sigma} \cdot \alpha_p$.
Zde α_p jsou kvantily normovaného normálního rozdělení definované aproximativně rov. (10) a $\hat{\mu}$, $\hat{\sigma}$ jsou vhodné odhady střední hodnoty resp. směrodatné odchylky. Při volbě $\hat{\mu} = \bar{x}$, $\hat{\sigma} = s$ (na grafech je označeno tečkovaně) lze sledovat také vhodnost použití předpokladu normality pro určení parametrů polohy. Pro případ, kdy se v datech vyskytují vybočující pozorování, je vhodné volit robustní odhady $\tilde{\mu} = \tilde{x}_{0.5}$ a $\tilde{\sigma} = \Delta_I/1.349$ (na grafech je označeno čárkovaně). Detaily o interpretaci těchto grafů uvádí práce /10, 11/.

Na obr. 1b je znázorněn kvantilový graf spolu s oběma průběhy kvantilové funkce normálního rozdělení.
Je patrná nesymetrie rozdělení (sešikmení vpravo) a dobrá aproximace normálním rozdělením (při použití robustních odhadů $\tilde{\mu}$, $\tilde{\sigma}$) až do kvartilů.

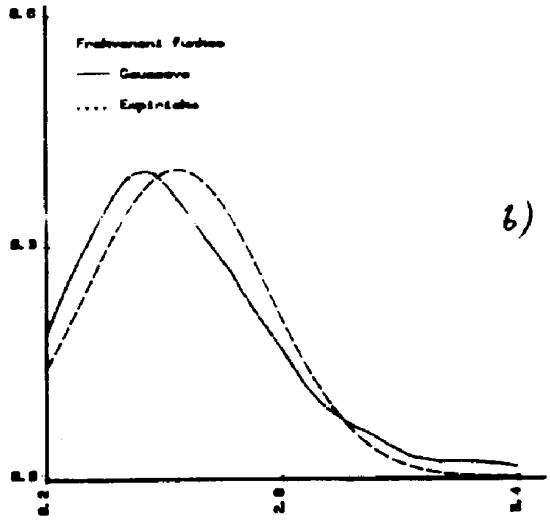
2.4 Grafy symetrie

Speciálně pro vyjádření stupně symetrie dat existuje řada variantních grafů /11/. Mezi nejjednodušší patří vynesení hodnot $a_i = (x_{(n+1-i)} - \tilde{x}_{0.5})$ proti hodnotám $b_i = (\tilde{x}_{0.5} - x_{(i)})$ $i = 1 \ldots [n/2]$. Pro symetrická data leží jednotlivé body v tomto grafu přibližně na přímce $a = b$. V okolí počátku se projevuje symetrie v oblasti mediánu a v okolí pravého horního rohu se projevuje symetrie v oblasti konců (detaily viz /10/).

Na obr. 2a je znázorněn graf symetrie spolu s teoretickou čarou $a = b$. Je patrná odchylka od symetrie v oblasti konců rozdělení dat.



a)



b)

Obr. 2 a) Graf symetrie
b) Frekvenční funkce.

2.5 Grafy empirické frekvenční funkce a histogramy

Řada technik z této skupiny se již řadí spíše mezi neparametrické statistické metody. Z různých variant (viz /10/) byla vybrána konstrukce empirické frekvenční funkce $\hat{f}(x)$ založená na "jádrovém" odhadu /11/

$$\hat{f}(x) = \frac{1}{N \cdot h} \sum_{i=1}^{N} \sigma\left[\frac{x - x_i}{h}\right]$$

kde jádrová funkce $\sigma[\cdot]$ je opět formálně frekvenční funkce symetrická kolem počátku a h je parametr vyhlazení. V programu je použita volba jádrové funkce v jednoduchém tvaru

$$\sigma(u) = \begin{cases} 0.75(1-u^2) & \text{pro } -1 \le u \le 1 \\ 0 & \text{jinde} \end{cases}$$

Parametr vyhlazení se standardně počítá podle vztahu $h = 2.35 \cdot s \cdot n^{-0.2}$

který vyhovuje pro nepříliš sešikmená rozdělení. Pro komplikovanější rozdělení se velikost h vhodně snižuje (viz /10/).

Na obr. 2b je uveden tento odhad frekvenční funkce tečkovaně. Pro porovnání je plně znázorněn průběh frekvenční funkce normálního rozdělení s parametry $\bar{x}$, s^2. Při konstrukci histogramu se využívá speciálního postupu volby šířky třídních intervalů, který je založen na odhadu výběrové kvantilové funkce (viz /10/).

3. Hodnocení výběrového rozdělení

V této skupině jsou zařazeny techniky, které umožňují odhad typu výběrového rozdělení resp. porovnání výběrového rozdělení se zvoleným teoretickým. Pro tyto účely jsou v programu "GR-POR" obsaženy tyto postupy:

1. Grafy rozptýlení s kvantily (GRK)
2. Kvantil-kvantilové grafy (Q-Q)
3. Empirické pravděpodobnostní grafy (EPP)
4. Stabilizované pravděpodobnostní grafy (SP)
5. Grafy transformované distribuční funkce (TDF)

Empirické rozdělení výběru je charakterizováno např. distribuční funkcí $F(x)$

a teoretické (předpokládané) rozdělení dat distribuční funkcí $F_T(x)$.
Pro grafy ze skupiny 2 až 5 je možno volit mezi 11ti teoretickými funkcemi $F_T(x)$.
Jsou to distribuční funkce pro tato rozdělení:

1. rovnoměrné	2. normální	3. lognormální	4. exponenciální
5. Laplaceovo	6. logistické	7. Paretovo +)	8. Cauchyho
9. Gumbellovo	10. Weibullovo +)	11. Gamma +)	

U rozdělení, označených +), je volitelný parametr tvaru.
Při konstrukci jednotlivých grafů se opět využívá vlastností pořádkových statistik $x_{(i)}$.

Poznámka: Uveďme si některé základní vlastnosti pořádkových statistik, které se uplatní při konstrukci jednotlivých grafů. Je-li $F_0(x)$ distribuční funkce rozdělení, ze kterého daný výběr pochází, má transformovaná náhodná veličina $z_{(i)} = F_0(x_{(i)})$ nezávisle na rozdělení $F_0(x)$ beta rozdělení $Be[i, n-i+1]$. Z toho plyne, že střední hodnota $E(z_{(i)})$ je rovna

$$E(z_{(i)}) = i/(n+1) \tag{1}$$

rozptyl $D(z_{(i)})$ je dán vztahem

$$D(z_{(i)}) = [i(n-i+1)] \cdot [(n+1)^2 \cdot (n+2)]^{-1} \tag{2}$$

a pro kovarianci mezi $z_{(i)}$ a $z_{(j)}$ platí

$$cov(z_{(i)}, z_{(j)}) = [i(n-j+1)] \cdot [(n+1)^2 \cdot (n+2)]^{-1} \tag{3}$$

Při zpětné transformaci $E(z_{(i)})$ na původní pořádkovou statistiku vyjde

$$E(x_{(i)}) = F_0^{-1}(E(z_{(i)})) = F_0^{-1}(P_i) \tag{4}$$

kde $P_i = i/(n+1)$ je tzv. pořadová pravděpodobnost. Z rov. (4) plyne, že $E(x_{(i)})$ je $100.P_i$%ní kvantil rozdělení $F_0(x)$. Pořádková statistika $x_{(i)}$ je pak odhad kvantilové funkce $Q_0(P_i) = F_0^{-1}(P_i)$ v místě P_i (u velkých výběrů lze tedy třídní hodnoty x_i^T brát jako odhady $Q_0(P_i)$ pro $P_i = \sum_{j=1}^{i} f_j - 0.5 f_i$ kde f_i jsou relativní četnosti).

Speciálně pro případ, že $F_0(x)$ je normální rozdělení, lze docílit zlepšení odhady $\hat{Q}_0(P_i)$ volbou

$$P_i = (i - 3/8) / (n + 0.25)$$

Pokud je třeba určit odhad $\hat{Q}(x)$ v místě $i/(n+1) < x < (i+1)/(n+1)$, je možno použít lineární interpolace ($\hat{Q}(x)$ je tedy uvažována jako lineární lomená čára)

$$\hat{Q}(x) = (n+1) \cdot (x - i/(n+1)) \cdot (x_{(i+1)} - x_{(i)}) + x_{(i)} \tag{5}$$

3.1 Graf rozptýlení s kvantily /2/

Tento graf umožňuje vyjádření dat ve vhodné formě, ve které lze analyzovat jejich statistické zvláštnosti. Využívá se odhadu kvantilové funkce $\hat{Q}(x)$ definovaného rov. (5). Na osu y se tedy vynáší pořádkové statistiky $x_{(i)}$ a na osu x odpovídající pořadové pravděpodobnosti P_i . Vzniklá soustava bodů se spojí lineárními úseky. Pro analýzu chování kvantilové funkce se do vzniklého grafu zakreslují tři obdélníky H, E a D.
Kvartilový obdélník H má x-ové souřadnice rovny $P_{H_1} = 0.25$ a $P_{H_2} = 0.75$. Odpovída-

jící y-ové souřadnice vrcholů jsou hodnoty kvartilů $\hat{Q}(P_{H1}) = \tilde{x}_{0.25}$ resp. $\hat{Q}(P_{H2}) = \tilde{x}_{0.75}$, které se určují lineární interpolací z rov. (5).

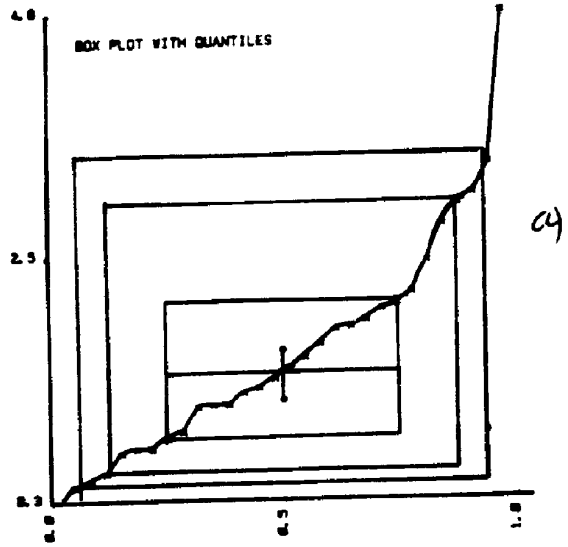
Oktilový obdélník E má x-ové souřadnice vrcholů $P_{E1} = 0.125$ a $P_{E2} = 0.875$ a y-ové souřadnice jsou oktily $\tilde{x}_{0.125}$ a $\tilde{x}_{0.875}$.

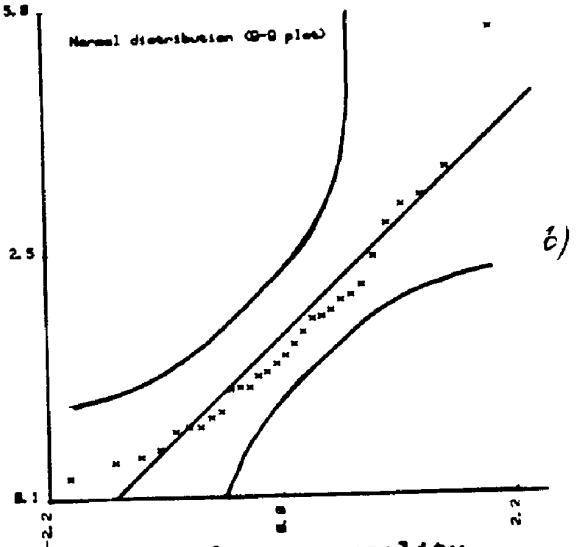
Sedecilový obdélník D má x-ové souřadnice vrcholů $P_{D1} = 0.0625$ a $P_{D2} = 0.9375$ a y-ové souřadnice jsou sedecily $\tilde{x}_{0.0625}$ resp. $\tilde{x}_{0.9375}$.

V kvartilovém obdélníku je ve výšce $\tilde{x}_{0.5}$ (medián) horizontální čára. Pro $P_i = 0.5$ (na ose x) se na kolmici k mediánové čáře vyznačuje robustní odhad konfidenčního intervalu

$$\tilde{x}_{0.5} \pm (\tilde{x}_{0.75} - \tilde{x}_{0.25}) / (n)^{1/2} \qquad (6)$$

GRK je pro zkoumaná data znázorněn na obr. 3a.





Obr. 3 a) GRK b) Q-Q graf pro normalitu.

Tento graf se ještě doplňuje tabelárními údaji pro $P = 0.25$ (kvartily, charakterizující chování výběrového rozdělení ve střední části), pro $P = 0.125$ (oktily, charakterizující chování dále od mediánu) a $P = 0.0625$ (sedecily, charakterizující chování v oblasti konců). Určují se tyto charakteristiky:

- <u>polosumy</u> $P_P = 0.5(\tilde{x}_P + \tilde{x}_{1-P})$. Pro symetrická rozdělení jsou P_P přibližně konstantní v celém rozmezí P.
- <u>rozpětí</u> $R_P = (\tilde{x}_{1-P} - \tilde{x}_P)$. Pro symetrická rozdělení musí R_P rovnoměrně růst s poklesem P.
- <u>relativní šikmosti</u> $S_P = (\tilde{x}_{0.5} - P_P)/R_P$. Pro symetrická rozdělení jsou S_P blízké nule.
- <u>délky konců</u> $T_P = Ln(R_P/R_{0.25})$. Délky konců $T_{0.125}$ a $T_{0.0625}$ jsou pro vybraná symetrická rozdělení uvedeny v tab. 1. Pozor, pro sešikmená rozdělení mohou vyjít délky konců i značně vyšší.

Tab. 1 Délky konců T_P (P = 0.125, 0.0625).

rozdělení	$T_{0.125}$	$T_{0.0625}$
normální	0.534	0.822
rovnoměrné	0.405	0.559
Laplaceovo	0.693	1.098

Pro data z obr. 3a jsou relativní šikmosti a délky konců uvedeny v tab. 2.

Tab. 2 Charakteristiky rozdělení dat z obr. 3a.

kvantil P	P_P	S_P	T_P
0.25	1.501	-0.025	0
0.125	1.796	-0.134	0.712
0.0625	1.958	-0.163	0.879

GRK, doplněný o výše uvedené charakteristiky, tedy umožňuje indikaci celé řady statistických zvláštností dat. Jsou to zejména:

- <u>symetrické smíšené rozdělení</u>, kdy jsou jednotlivé obdélníky H, E, D v sobě symetricky umístěny. Navíc jsou P_P přibližně konstantní a $S_P \sim 0$. Pokud jsou navíc délky konců přibližně rovny délkám konců z tab. 1, lze identifikovat <u>normální</u>, <u>rekvangulární</u> resp. <u>Laplaceovo</u> rozdělení,
- <u>nesymetrické rozdělení</u>, kdy jsou vzdálenosti mezi spodními a horními hranami obdélníků H, E, D různé. Pro sešikmení k vyšším hodnotám jsou vzdálenosti mezi spodními hranami H, E, D kratší než mezi horními hranami. Navíc P_P rostou s poklesem P a S_P jsou záporné (v absolutní hodnotě rostou s poklesem P). U rozdělení sešikmených k nižším hodnotám je tomu naopak.
- <u>náhlé pozorování</u>, kdy je na $\hat{Q}(P)$ mimo obdélník H náhlý vzrůst (směrnice $\hat{Q}(P)$ roste prakticky nade všechny meze).
- <u>bimodální rozdělení</u>, kdy je na $\hat{Q}(P)$ náhlý vzrůst uvnitř obdélníku H.

Pro zpracovávaný výběr lze tedy na základě tab. 2 a obr. 3a určit, že rozdělení dat je v oblasti konců sešikmené směrem k vyšším hodnotám a navíc je zřejmě přítomno jedno vybočující pozorování.

3.2 Kvantilo-kvantilové grafy

Tyto grafy (označované jako Q-Q) slouží k posouzení shody výběrového rozdělení, charakterizovaného kvantilovou funkcí $\hat{Q}(u)$, se zvoleným teoretickým rozdělením, charakterizovaného kvantilovou funkcí $Q_T(u) = F_T^{-1}(u)$.
Vychází se přitom z faktu, že v případě shody obou rozdělení, jsou shodné obě kvantilové funkce.
Jak již bylo uvedeno, jsou hrubým odhadem výběrové kvantilové funkce $\hat{Q}(u)$ přímo pořádkové statistiky $x_{(i)}$. Teoretické kvantilové funkce $Q_T(u)$ závisí také na hodnotách jistých parametrů, pro které nejsou ve fázi průzkumové analýzy k dispozici ani odhady. Pro teoretické distribuční funkce typu $F_T((x-a)/b)$ lze standardizací $S=(x-a)/b$ dospět ke standardizovaným distribučním funkcím $F_{ST}(S)$. Odpovídající standardizované kvantilové funkce $Q_{ST}(u)$ již neobsahují parametry a , b .

<u>Poznámka</u>: Parametry a resp. b mají obvykle význam <u>polohy</u> resp. <u>měřítka</u>. Pokud obsahuje rozdělení F_T ještě další parametry tvaru (Weibullovo, Gamma, Paretovo), počítají se $Q_{ST}(u)$ pro jejich zvolené hodnoty. Pro většinu rozdělení lze $Q_{ST}(u)$ vyjádřit analytickým výrazem. Pro $Q_{ST}(u)$ normálního rozdělení se používá jednoduchá aproximace definovaná rov. (10). Také pro Gamma rozdělení je nalezena jednoduchá analytická aproximace /10/.

<u>Q-Q grafy</u> jsou závislosti pořádkových statistik $x_{(i)}$ na odpovídajících standardizovaných kvantilech $Q_{ST}(P_i)$ zvoleného teoretického rozdělení (běžně

se opět sosazuje $P_i = i/(n+1)$).
V případě shody výběrového rozdělení s teoretickým vznikne v Q-Q grafu lineární závislost typu

$$x_{(i)} = a + b \cdot Q_{ST}(P_i) \qquad (7)$$

Q-Q grafy, konstruované v programu "GR-POR", obsahují kromě bodů $[x_{(i)}, Q_{ST}(P_i)]$ také odpovídající regresní přímku, určenou metodou nejmenších čtverců. Pro zhodnocení variability dat v Q-Q grafech lze konstruovat buď asymptotické konfidenční pásy nebo směrodatné odchylky pro jednotlivé body (viz /10, 11/).

Pro asymptotické konfidenční pásy (odpovídající Kolmogorovu testu shody) platí

$$Q_T(P_i - C_{1-\alpha}) \leq x_{(i)} \leq Q_T(P_i + C_{1-\alpha}) \qquad (8)$$

Zde $C_{1-\alpha}$ je parametr, závislý na hladině významnosti α a na rozsahu výběru n. Pro standardní 95%ní konfidenční pásy je $C_{1-\alpha}$ vyjádřitelné aproximativním vztahem /3/

$$C_{0,95} = 1{,}385\left[n + \sqrt{n+4}\,/3{,}5\right]^{-1/2} \qquad (9)$$

V prvních fázích průzkumové analýzy se používají Q-Q grafy pro ověření normality výběrového rozdělení, které se běžně označuje jako <u>rankitové</u>.

<u>Poznámka</u>: Pro určení kvantilu $Q_{ST}(P_i)$ normovaného normálního rozdělení se využívá jednoduché aproximace

$$Q_{ST}(P_i) \equiv z_p = \frac{-9{,}4 \cdot \ln(1/P_i - 1)}{\sqrt{\ln[\ln(1/P_i - 1)]} + 14} \qquad (10)$$

Tyto grafy mohou navíc sloužit pro orientační zařazení výběrového rozdělení do skupin dle šikmosti a špičatosti (délky konců). Pro rankitové grafy platí, že:

- <u>konvexně-konkávní</u> průběh — indikuje rozdělení s krátkými konci
- <u>konkávně-konvexní</u> průběh — indikuje rozdělení s dlouhými konci
- <u>konkávní průběh</u> — indikuje rozdělení sešikmené vlevo
- <u>konvexní průběh</u> — indikuje rozdělení sešikmené vpravo

Pro určení míry shody mezi $x_{(i)}$ a jednotlivými $Q_{ST}(P_i)$ se kromě vizuální inspekce počítají také korelační koeficienty.

Na obr. 3b je uveden Q-Q graf pro ověření normality (rankitový graf). Je zřejmé, že rozdělení zpracovávaného výběru je sešikmené vpravo. Navíc jsou asymptotické konfidenční pásy příliš "široké".

3.3 Empirické pravděpodobnostní grafy

Tyto grafy jsou obdobou Q-Q grafů. Místo teoretických kvantilů $Q_T(P_i)$ se počítají na základě simulovaných výběrů (ze zvoleného $F(x)$) speciální kvantily T_i. Stejným způsobem se tvoří konfidenční pásy (analogický postup pro případ logistické regrese je popsán v /4/).

Určení kvantilů T_i a koncových bodů LB_i, UB_i pro 85%ní konfidenční pásy lze sumarizovat do těchto kroků:

1. Volba $F_T(x)$ a určení maximálně věrohodných odhadů jeho parametrů na základě výběru.

2. Konstrukce simulovaných výběrů $\{x_i^{(j)}\}$ i=1,...n, j=1,...25 z rozdělení $F_T(x)$ (např. pro posouzení normality se vybírá z rozdělení $N(\bar{x}, \hat{S}^2)$).

3. Tvorba pořádkových statistik (uspořádání) $\{x_{(i)}^{(j)}\}$ i=1,...n pro každý simulovaný výběr j=1,...25 zvlášť.

4. Určení "simulačních" kvantilů T_i jako mediánů z výběrů j=1,...25, tedy

$$T_i = med\{x_{(i)}^{(1)}, \ldots x_{(i)}^{(25)}\} \qquad i = 1, \ldots n$$

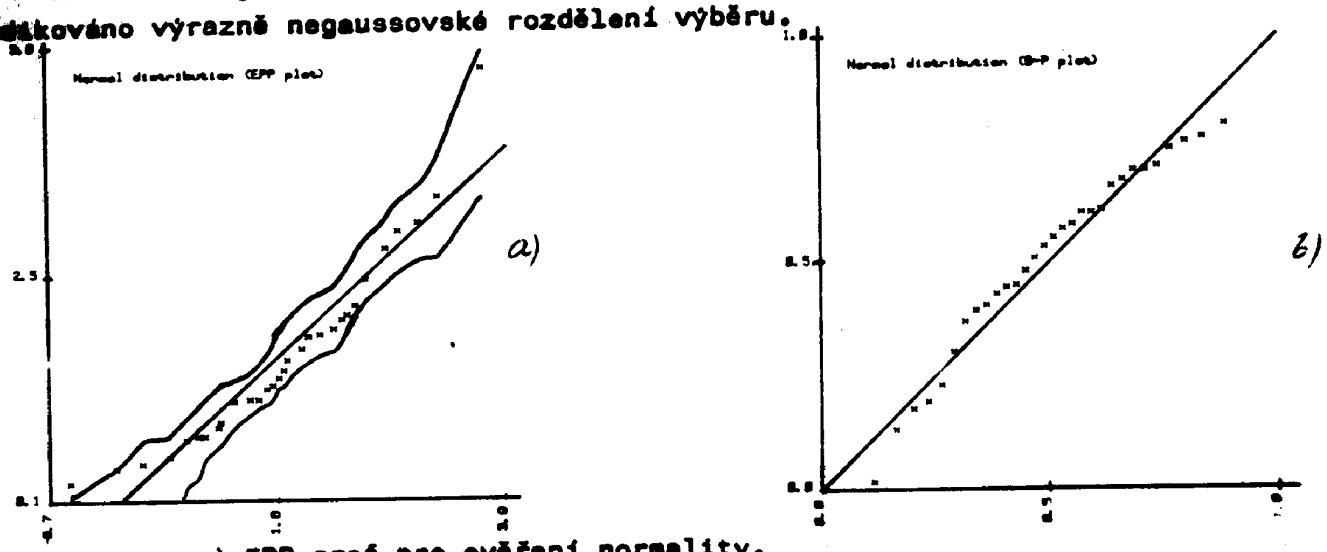
5. Určení spodní LB_i a horní UB_i meze 85%ní konfidenční oblasti

LB_i = druhá nejmenší hodnota v $\{x_{(i)}^{(1)}, \ldots x_{(i)}^{(25)}\}$

UB_i = druhá největší hodnota v $\{x_{(i)}^{(1)}, \ldots x_{(i)}^{(25)}\}$ i=1,...n

Platí, že pokud $F_Q(x) \sim F_T(x)$, je $x_{(i)} = T_i$. V EPP grafu, kde se vynáší $x_{(i)}$ vs. T_i , budou tedy jednotlivé body ležet na přímce se směrnicí 1 a úsekem 0 . EPP graf je doplněn o regresní přímku, určenou metodou nejmenších čtverců.

Na obr. 4a je uveden EPP graf pro ověření normality. Je zřejmé, že je opět indikováno výrazně negaussovské rozdělení výběru.



Obr. 4 a) EPP graf pro ověření normality.
b) SPP graf pro ověření normality.

3.4 Stabilizované pravděpodobnostní grafy

Také S-P grafy se používají ke stejným účelům jako Q-Q resp. EPP grafy. Jde vlastně o speciálně upravené pravděpodobnostní (P-P) grafy tak, aby bylo dosaženo stabilizace rozptylu /5/.

U klasických P-P grafů se na osu y vynášejí pravděpodobnosti $u_{(i)} = F_T(x_{(i)}, \hat{a})$ kde $\hat{a}$ jsou odhady parametrů daného rozdělení F_T získané z výběru metodou maximální věrohodnosti. Na osu x se vynášejí pořadové pravděpodobnosti $P_i = \frac{i}{n+1}$. Pokud $F_Q(x) \sim F_T(x)$, mají $u_{(i)}$ rovnoměrné rozdělení (jde o transformaci náhodné veličiny vzhledem k její distribuční funkci).

Poznámka: Transformace $S = (2/\pi) \cdot \arcsin(u)^{1/2}$, kde u má rovnoměrné rozdělení R(0,1), vede na sinové rozdělení s frekvenční funkcí

$$0{,}5 \cdot \pi \cdot \sin(\pi S) \qquad (0 < S < 1)$$

Pořádkové statistiky $S_{(i)}$ tohoto rozdělení mají asymptoticky stejné rozptyly /5/.

U S-P grafu se na osu y vynášejí $S_{(i)} = (2/\pi) \arcsin(u_{(i)})^{1/2}$ a na osu x hodnoty $T_{(i)} = (2/\pi) \arcsin(P_i)^{1/2}$. Pokud platí, že $F_Q(x) \sim F_T(x)$, je $S_{(i)} = T_{(i)}$. Jednotlivé body v S-P grafu pak leží na přímce se směrnicí 1 a nulovým úsekem. Na obr. 4b je uveden S-P graf pro ověření normality rozdělení.

zpracovávaného výběru. Indikuje prakticky totéž co EPP a Q-Q graf.

3.5 Grafy transformované distribuční funkce /2/

Grafy Q-Q, EPP i S-P pro porovnání shody $F_e(x)$ a $F_T(x)$ mají dvě základní nevýhody:

- o linearitě se usuzuje pouze z n bodů (což je pro malá n nepřesné)
- u EPP a S-P grafů je třeba počítat maximálně věrohodné odhady parametrů pro dané $F_T(x)$.

Odstranění obou těchto nevýhod lze docílit použitím TDF grafů, které vycházejí z faktu, že pokud $F_e(x) \sim F_T(x)$, platí

$$f_e[Q_e(P)] = \frac{1}{\sigma} \cdot f_T[Q_T(P)] \qquad (11)$$

kde σ je parametr měřítka a symbol $f[Q(P)]$ označuje frekvenční kvantilovou funkci.

Poznámka: Uvědme, že pro distribuční funkci $F(x)$ a frekvenční funkci $f(x) = F'(x)$ resp. kvantilovou funkci $Q(P) = F^{-1}(x)$ lze definovat:

- frekvenční kvantilovou funkci $f[Q(P)]$
- kvantilovou frekvenční funkci $q(P) = Q'(P)$.

Přitom platí rovnost

$$f[Q(P)] \cdot q(P) = 1 \qquad (12)$$

Z rovnosti (12) plyne, že pro $F_e(x) \sim F_T(x)$ bude tzv. <u>transformované frekvenční funkce</u>

$$d(P) = \frac{1}{\sigma_T} \cdot f_T[Q_T(P)] \cdot q_e(P) \qquad (13)$$

rovna 1 pro $0 \le P \le 1$. V rov. (13) má σ_T význam normalizační konstanty

$$\sigma_T = \int_0^1 f_T[Q_T(z)] \cdot q_e(z) \cdot dz \qquad (14)$$

Podobně bude transformovaná distribuční funkce

$$D(P) = \int_0^P d(z) \cdot dz \qquad (15)$$

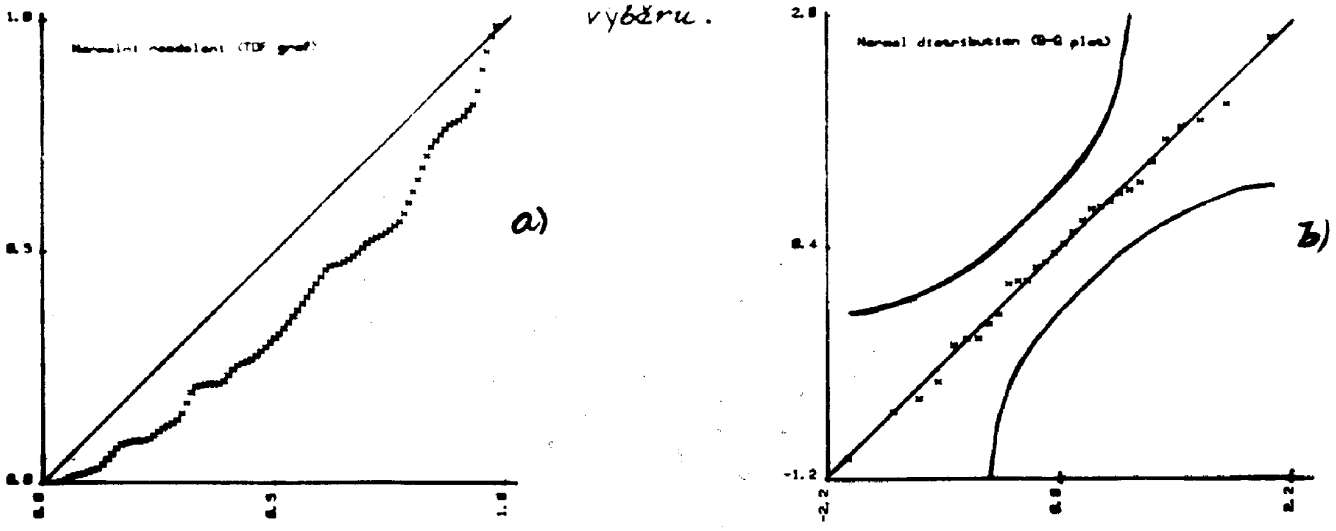
rovna P .

Odchylky $D(P)$ od přímky $D(P)=P$ tedy indikují míru neshody $F_e(x)$ a $F_T(x)$. V TDF grafu se tedy vynášejí hodnoty $D(P)$ vs. P (pro zvolené dělení P). Kvantilová frekvenční funkce $q_e(P)$ se nahrazuje odhadem

$$\hat{q}(P) = n \cdot [x_{(j+1)} - x_{(j)}] \qquad \text{pro} \quad (2j-1)/n < P < (2j+1)/n \qquad (16)$$

Rov. (16) vychází z předpokladu, že je $\hat{Q}(P)$ aproximována lineárními lomenými úseky. Pro dané teoretické rozdělení $F_T(x)$ se dá $f_T[Q_T(P)]$ určit pouhým dosazením. Např. pro $F_T = 1-\exp(-x)$ je $Q_T(P) = -\ln(1-P)$ a $f_T(x) = \exp(-x)$. Pak $f_T[Q_T(P)] = 1-P$. Hodnoty $D(P)$ z rov. (15) se poměrně snadno určí numerickou integrací.

Na obr. 5a je vynesen TDF graf pro posouzení normality rozdělení zpracovávaného výběru.



Obr. 5 a) TDF graf pro posouzení normality
b) Q-Q graf pro ověření normality v Box-Coxově transformaci (λ =0.2).

Porovnáme-li obr. 3 až 5, zjistíme, že TDF (obr. 5a) je nejcitlivější na odchylky výběrového rozdělení od normality. To platí i pro další teoretická rozdělení.

4. Mocninné transformace

Mocninná transformace se poněkud vymyká z postupů průzkumové analýzy, protože kromě "exploratorní" identifikace odchylek od normality nabízí alternativu, jak "zlepšit" rozdělení dat. Zde se omezíme na rodiny mocninných transformací, které obyčejně vedou k:

- odstranění (mocninné) závislosti mezi střední hodnotou a rozptylem,
- zesymetričtění výběrového rozdělení,
- stabilizace rozptylu,
- přiblížení výběrového rozdělení k normálnímu.

Ne vždy vede pochopitelně mocninná transformace k statisticky významnému zlepšení rozdělení výběru. Vznikají zde také problémy s její interpretací a další statistickou analýzou transformovaných dat. Přesto je často nejjednodušší možností jak umožnit zpracování dat klasickými postupy (vycházejícími z předpokladu normality).

Poznámka: Nejjednodušší je tzv. prostá mocninná transformace, pro kterou platí

$$x^{(\lambda)} = x^{\lambda} \quad (\lambda > 0) \qquad x^{(\lambda)} = \ln x \quad (\lambda = 0) \qquad x^{(\lambda)} = -x^{\lambda} \quad (\lambda < 0)$$

Zde již zřejmě není aritmetický průměr vždy odhadem střední hodnoty. Pro λ =-1 to bude harmonický průměr, pro λ =0 geometrický průměr a pro λ =2 kvadratický.

Nejznámější rodinou mocninných transformací je Box-Coxova transformace, pro kterou platí /6/

$$x^{(\lambda)} = (x^{\lambda} - 1)/\lambda \qquad \text{pro} \quad \lambda \neq 0$$
$$x^{(\lambda)} = \lim_{\lambda \to 0} [(x^{\lambda} - 1)/\lambda] = \ln x \qquad \text{pro} \quad \lambda = 0 \qquad (17)$$

Je zřejmé, že takto definovaná transformace je spojitá pro všechna λ . Má navíc tyto základní vlastnosti

$x^{(\lambda)} > 0$ pro všechna $x > 0$
$x^{(\lambda)} \leq 0$ pro všechna $0 < x \leq 1$
$x^{(\lambda)}$ je rostoucí funkcí x i λ
$x^{(\lambda)}$ je konvexní vzhledem k λ pro $x \geq 1$ a
je konkávní vzhledem k λ pro $x < 1$.

Předpokladem pro použití transformace (17) je, že $\{x_i\}$ i=1,...n jsou realizace kladné náhodné veličiny (což také koliduje s předpokladem, že tato transformace vede k normalitě).

Pro případ, že ve výběru jsou i záporné hodnoty, používá se <u>zobecněné</u> Box-Coxovy transformace.

Pokud je výběrové rozdělení sice symetrické, ale má v porovnání s normálním rozdělením dlouhé konce (je špičatější), lze docílit přibližné normality <u>modulovou</u> transformací /7/.

Emerson a Stotto /8/ navrhli pro účely průzkumové analýzy rodinu transformací $x^{(\lambda)} = A_\lambda \cdot x^\lambda + B_\lambda$ (pro $\lambda \neq 0$) resp. $x^{(\lambda)} = A_0 \cdot \ln x + B_0$ (18) (pro $\lambda = 0$).

Parametry A_λ, B_λ se volí tak, aby docházelo při transformaci k minimálním změnám v okolí mediánu, tedy

$$\text{med}(x^{(\lambda)}) = \text{med}(x) \qquad \frac{d}{dx}\,\text{med}(x^{(\lambda)}) = 1 \qquad (19)$$

Pro odhad vhodného parametru λ v těchto rodinách mocninných transformací je možno použít různých technik.

<u>a) Empirická kritéria</u>. Do této skupiny patří řada kritérií definujících tvar transformovaného rozdělení. Základní jsou:

a1) šikmost S_K (klasické definice), kdy se za optimální považuje λ, pro které má výběr $\{x_i^{(\lambda)}\}$ i=1,...n minimální absolutní hodnotu S_K.

a2) špičatost K_u-3 (klasická definice). Opět je optimální to λ, které vede k minimu absolutní hodnoty této charakteristiky.

a3) asymetrie $$A = [\bar{x}^{(\lambda)} - \tilde{x}_{0.5}^{(\lambda)}] / (x_{(n)}^{(\lambda)} - x_{(1)}^{(\lambda)}) \qquad (20)$$

kde $\bar{x}^{(\lambda)}$ resp. $\tilde{x}_{0.5}^{(\lambda)}$ je aritmetický průměr resp. medián transformovaných dat. Za optimální se považuje λ vedoucí k minimu abs(A). Místo A je možné použít robustní asymetrii A_R definovanou podle vztahu

$$A_R = \{[\tilde{x}_{0.75}^{(\lambda)} - \tilde{x}_{0.5}^{(\lambda)}] - [\tilde{x}_{0.5}^{(\lambda)} - \tilde{x}_{0.25}^{(\lambda)}]\} / (\tilde{x}_{0.75}^{(\lambda)} - \tilde{x}_{0.25}^{(\lambda)}) \qquad (21)$$

kde $\tilde{x}_{0.75}^{(\lambda)}$ resp. $\tilde{x}_{0.25}^{(\lambda)}$ je horní resp. spodní kvartil transformovaných dat.

<u>b) Statistické postupy</u>. Nejpoužívanější v této skupině je metoda maximální věrohodnosti. Vychází se z předpokladu, že $x_i^{(\lambda)}$ jsou nezávislé náhodné veličiny s normálním rozdělením $N(\bar{x}^{(\lambda)}, \sigma^2(\lambda))$. Odpovídající frekvenční funkce pro x_i má pak tvar

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_{(x)}^2}} \cdot \exp\left[\frac{-1}{2\sigma_{(x)}^2} \cdot \sum_{(i)} (x_i^{(\lambda)} - \bar{x}^{(\lambda)})^2\right] \cdot \left|\frac{d}{dx}(x_i)^{(\lambda)}\right| \qquad (22)$$

<u>Poznámka</u>: Pokud se provede normalizační transformace Z_i taková, že $\frac{d}{dx}(x_i^{(\lambda)}) = 1$ vypadne v rov. (22) člen v absolutní hodnotě.

Nahradíme-li v rov. (22) parametry $\bar{x}^{(\lambda)}$, $\sigma^2(\lambda)$ jejich maximálně věrohodnými odhady $\hat{\bar{x}}^{(\lambda)} = n^{-1} \sum_{(i)} x_i^{(\lambda)}$; $\hat{\sigma}^2_{(\lambda)} \sim n^{-1} \sum_{(i)} (x_i^{(\lambda)} - \bar{x}^{(\lambda)})^2$ můžeme snadno konstruovat věrohodnostní funkci $\ln L(\lambda)$ závislou pouze na λ.

Pro případ Box-Coxovy transformace je navíc $\frac{d}{dx}(x_i^{(\lambda)}) = x_i^{\lambda-1}$, takže dospíváme ke známému kritériu

$$\ln L_B(\lambda) = -\frac{n}{2} \cdot \ln(\hat{\sigma}^2_{(\lambda)}) + (\lambda - 1) \sum_{(i)} \ln |x_i| \qquad (23)$$

Maximalizací $\ln L_B(\lambda)$ (např. v intervalu $-3 < \lambda \leq 3$) lze určit maximálně věrohodný odhad $\hat{\lambda}$ a odpovídající $100(1-\alpha)\%$ní konfidenční interval

$$2[\ln L_B(\hat{\lambda}) - \ln L_B(\lambda)] \leq \chi^2_{1-\alpha}(1) \qquad (24)$$

kde $\chi^2_{1-\alpha}(1)$ je kvantil χ^2-rozdělení s jedním stupněm volnosti. S výhodou lze využít grafického znázornění $\ln L_B(\lambda)$ vs. λ (pokud v oblasti vymezené konfidenčním intervalem (24) leží $\lambda = 1$, nebude mít zřejmě mocninné transformace významný vliv na zlepšení rozdělení dat).

Kromě metody maximální věrohodnosti lze využít i dalších statistických metod (jako např. maximalizace á posteriorní hustoty pravděpodobnosti).

e) Exploratorní techniky. Speciálně pro rodinu transformací definovanou rov. (18) lze odhadnout optimální λ z grafu, kde se na osu y vynášejí souřadnice $y_i^* = (\tilde{x}_q - \tilde{x}_{1-q})/2 - \tilde{x}_{0.5}$ a na osu x hodnoty $x_i^* = [(\tilde{x}_{1-q} - \tilde{x})^2 + (\tilde{x} - \tilde{x}_q)^2]/4\tilde{x}_{0.5}$ pro $q = 1/2^i$, $i = 2, 3 \ldots$ ($\tilde{x}_q$ je q-kvantil). Pokud existuje mocninná transformace, vedoucí k zesymetričtění výběrového rozdělení, leží body $\{y_i^*, x_i^*\}$ na přímce se směrnicí $(1-\lambda)$.

Poznámka: Je zřejmé, že pro symetrická výběrová rozdělení jsou $y_i^* = 0$ pro všechna i a optimální $\lambda = 1$.

Program "MOC-TR" umožňuje nalezení vhodného λ pro transformaci, definovanou rov. (21) resp. (17). V prvním případě (rov. (21)) se λ stanovuje přímo ze speciálního grafu (závislost y^* na x^*). Ve druhém případě (rov. (17)) se pro určení λ používá různých kritérií. Optimální λ se určuje systematickým krokovým hledáním na intervalu $-3 \leq \lambda \leq 3$.
Pro ověření zlepšení rozdělení dat po mocninné transformaci se užívá všech typů grafů z kap. 3. V tab. 3 jsou uvedeny optimální hodnoty λ určené podle výše uvedených kritérií pro zpracování dat.

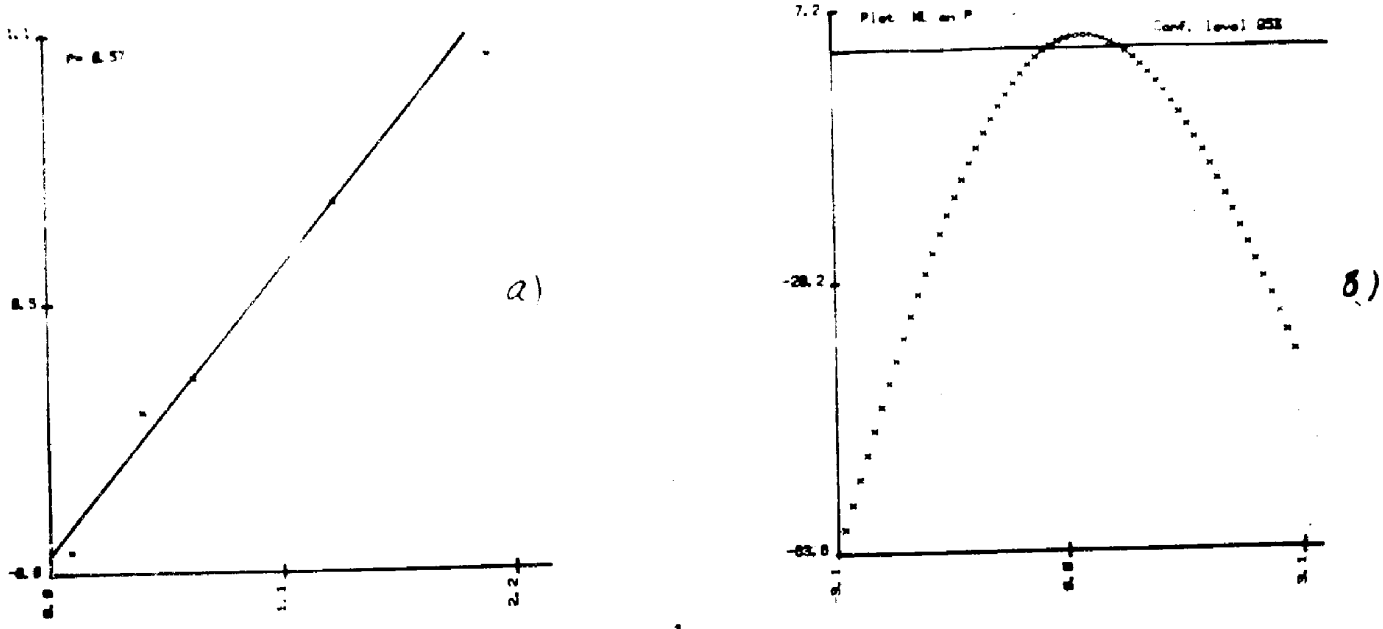
Tab. 3 Optimální λ dle různých kritérií.

kritérium	S_K	$K_U - 3$	A	A_R	$\ln L_B(\lambda)$	ES+)
$\hat{\lambda}$	0.2	-0.2	0.3	0.8	0.2	0.43
hodnota kritéria	0.066	0.01	0.0031	0.0063	3.96	-

+) určeno z grafu 6b.
Na obr. 6a je znázorněna závislost věrohodnostní funkce $\ln L_B(\lambda)$ na parametru λ. Je patrné výrazné maximum a poměrně úzký konfidenční interval pro $\hat{\lambda} = 0.2$.

Na obr. 5b je zakreslen Q-Q graf pro ověření normality výběru v Box-Coxově transformaci (rov.(17)) při volbě $\lambda = 0.2$. Je zřejmé, že mocninná transformace výrazně přispěla ke zlepšení rozdělení dat (přiblížení k normalitě). Navíc se zde žádná hodnota nejeví jako silně vybočující.

Na obr. 6b je uvedena závislost y_i^* na x_i^* pro grafický odhad λ dle Emerona a Stotta. Opět je patrná výrazná lineární závislost indikující vhodnost mocninné transformace.



Obr. 6 a) Závislost $\ln L_8(\lambda)$ na λ
b) Určení optimálního λ dle Emerson-Stotta.

5. Závěr

Programový soubor EXDAB umožňuje snadnou realizaci metod průzkumové analýzy dat. Je vhodný pro interaktivní práci s výpočetním prostředkem, který má možnost grafického výstupu.

Literatura

/1/ Tukey J.W.: Exploratory Data Analysis, Addison Wesley, Reading Inc. 1977 (překlad v ruštině Moskva 1982)

/2/ Parzen E.: J.A.S.A. 74, 105 (1979)

/3/ Harter L.: Amer. Statist. 34, 110 (1980)

/4/ Landwehr J.M., Pregibon D., Shoemaker A.C.: J.A.S.A. 79, 61 (1984)

/5/ Michael J.R.: Biometrika 70, 11 (1983)

/6/ Box G.E.P., Cox D.R.: J.R.Stat.Soc. B26, 211 (1964)

/7/ Box G.E.P., John E.T.: Appl. Statist. 29, 190 (1980)

/8/ Emerson J.D., Stotto M.A.: J.A.S.A. 77, 103 (1982)

/9/ Hinkley D.V.: Appl. Statist. 26, 67 (1977)

/10/ Militký J., Militká D.: Moderní matematicko-statistické metody v hutnictví, díl III, skripta ZP ČSVTS NHKG Ostrava 1985

/11/ Chambers J. a kol.: Graphical Methods for Data Analysis, Duxbury Press, Boston 1983