VĚROHODNOST VÝSLEDKŮ PŘI UŽITÍ EXPLORATORNÍ ANALÝZY DAT

Milan Meloun

Univerzita Pardubice, Čs. Legií 565, 532 10 Pardubice, milan.meloun@upce.cz

1. Obecný postup analýzy jednorozměrných dat

V prvním kroku se v exploratorní analýze dat vyšetřují *statistické zvláštnosti*, jako je lokální koncentrace dat, tvarové zvláštnosti rozdělení dat a přítomnosti podezřelých hodnot. Odhalí se tak anomálie a odchylky rozdělení výběru od předpokládaného symetrického rozdělení Gaussova, (obr. 1). Příkladem asymetrického rozdělení je log-normální rozdělení (obr. 2).

V druhém kroku se ověří *základní předpoklady*, kladené na výběr, jako jsou nezávislost prvků, homogenita výběru, dostatečný rozsah výběru a rozdělení výběru. Jsou-li závěry tohoto kroku optimistické, následuje vyčíslení klasických odhadů polohy a rozptýlení, tj. aritmetického průměru a rozptylu ve kroku čtvrtém. Sem patří i konstrukce intervalů spolehlivosti a případně testování hypotéz. V opačném případě následuje pokus o třetí krok, obsahující symetrizující transformaci dat.

Ve třetím kroku se provádí mocninná a Boxova-Coxova transformace, které mohou vést k symetričtějšímu rozdělení výběru a umožňují provedení korektnějšího odhadu polohy a rozptýlení. Transformace je vhodná především při nekontantnosti rozptylu a při asymetrii rozdělení původních dat.

Ve čtvrtém kroku se v konfirmatorní analýze nabízí paleta rozličných odhadů polohy, rozptýlení a tvaru, jež lze rozdělit do skupin: *klasické odhady* a *robustní odhady* (necitlivé na odlehlé prvky výběru, resp. další předpoklady o datech). Z nabídky odhadů parametrů vybírá uživatel uvážlivě ty, jež mají statistický smysl a odpovídají závěrům průzkumové analýzy dat a závěrům ověření předpokladů o výběru.

2. Průzkumová analýza dat (EDA)

2.1 Diagnostické grafy

Prvním krokem v analýze jednorozměrných dat je průzkumová analýza. Vychází se z *pořádkových statistik* výběru tj. z prvků výběru, uspořádaných vzestupně $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. Platí, že střední hodnota *i*-té pořádkové statistiky $x_{(i)}$ je rovna $100P_i$ procentnímu kvantilu výběrového rozdělení $Q(P_i)$ a symbol $P_i \approx i/(n+1)$ označuje *pořadovou pravděpodobnost*. V průzkumové analýze se často používá speciálních kvantilů L pro pořadové pravděpodobnosti $P_i = 2^{-i}$, $i = 1, 2, \ldots$, které se nazývají *písmenové hodnoty*. Kromě mediánu existují vždy dvojice kvantilů, definujících dolní a horní písmenové hodnoty L_D a L_H, např. F_D pro $P_i = 2^{-2}$ a F_H pro $P_i = 1 - 2^{-2}$.

i	i-tý kvantil	Pořadová pravděpodobnost P_i	Písmenová hodnota 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*

Počet prakticky určitelných písmenových hodnot ve výběru velikosti n_L závisí na jeho rozsahu n, a to dle vzorce $n_L \approx 1.44 \ln(n+1)$.

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

Rozmítnutý diagram rozptýlení (osa *x*: hodnoty *x*, osa *y*: interval náhodných čísel). Diagram představuje projekci kvantilového grafu, body jsou však vhodně rozmítnuté ve směru osy *y*.

Krabicový graf (osa *x*: úměrná hodnotám *x*, osa *y*: libovolný interval). Pro částečnou sumarizaci dat lze využít krabicového grafu, který umožňuje znázornění robustního odhadu polohy, mediánu *M*, dále posouzení symetrie v okolí kvartilů, posouzení symetrie u konců rozdělení, a konečně identifikaci odlehlých dat. Krabicový graf je obdélník o délce $R_F = F_H - F_D$ s vhodně zvolenou šířkou, která je úměrná hodnotě $\sqrt{n}$. V místě mediánu *M* je vertikální čára a úsečky obou krabicových fousů jsou ukončeny *vnitřními hradbami* B_H, B_D,dle

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

Prvky výběru mimo interval vnitřních hradeb [B_H, B_D,] jsou považovány za podezřelé (obvykle vybočující body), což je v grafu znázorněno kroužky.

Kvantilový graf (osa *x*: pořadová pravděpodobnost P_i, osa *y*: pořádková statistika $x_{(i)}$). Umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, který může být symetrický, sešikmený k vyšším nebo nižším hodnotám. Ke snadnějšímu porovnání s normálním rozdělením se do tohoto grafu zakreslují i kvantilové funkce, $N_{P_i} = \hat{\mu} + \hat{\sigma}\ u_{P_i}$, pro $0 \le P_i \le 1$: (1) *klasických odhadů* parametrů polohy a rozptýlení, tj. aritmetického průměru a směrodatné odchylky $\hat{\mu} = \bar{x}$ 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 interkvartilové rozpětí.

Graf polosum (osa *x*: pořádkové statistiky $x_{(i)}$, osa *y*: $Z_i = 0.5\ (x_{(n+1-i)} + x_{(i)})$). Pro symetrické rozdělení je grafem horizontální přímka, určená rovnicí $y = M$ a body oscilují okolo ní a vykazují náhodný shluk (mrak). Asymetrické rozdělení vykazuje nenáhodný trend.

Graf symetrie (osa *x*: $M - x_{(i)}$, osa *y:* $x_{(n+1-i)} - M$). Pro případ symetrického rozdělení resultuje lineární závislost s nulovým úsekem a jednotkovou směrnicí.

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

Pro normální rozdělení je grafem horizontální přímka s nulovou směrnicí. Pokud body leží na přímce s nenulovou směrnicí je tato směrnice odhadem parametru špičatosti.

Graf rozptýlení s kvantily (osa *x*: P_i, osa *y*: $x_{(i)}$). Základem je odhad kvantilové funkce výběru, který se získá spojením bodů $\{x_{(i)}, P_i\}$ lineárními úseky. Pro symetrická rozdělení má kvantilová funkce sigmoidální tvar. Pro rozdělení sešikmená k vyšším hodnotám je konvexně rostoucí a pro rozdělení sešikmená k nižším hodnotám konkávně rostoucí. Do grafu se zakreslují tři obdélníky: *kvartilový obdélník F:* na *y* ose kvartily F_D a F_H a na ose *x* pořadové pravděpodobnosti $P_2 = 2^{-2} = 0.25$ a $1 - 2^{-2} = 0.75$ a analogicky další dva obdélníky, *oktilový obdélník E* a *sedecilový obdélník D*. Graf umožnuje určení diagnóz: 1. *Symetrické unimodální rozdělení výběru* obsahuje vždy obdélníky symetricky uvnitř sebe. 2. *Nesymetrická rozdělení* mají pro rozdělení sešikmené k vyšším hodnotám vzdálenosti mezi dolními hranami obdélníků *F, E* a *D* výrazně kratší než mezi jejich horními hranami. 3. *Odlehlá pozorování* jsou indikována tím, že na kvantilové funkci mimo obdélník *D* se objeví náhlý vzrůst, kdy hodnota směrnice roste nade všechny meze. 4. *Vícemodální rozdělení* jsou indikována tím, že na kvantilové funkci uvnitř obdélníku *F* je několik úseků s téměř nulovými směrnicemi.

Histogram (osa *x*: proměnná *x*, osa *y*: uměrná hustotě pravděpodobnosti). Jde o obrys sloupcového grafu, kde jsou na ose *x* jednotlivé třídy, definující šířky sloupců, a výšky sloupců odpovídají empirickým hustotám pravděpodobnosti. Kvalitu histogramu ovlivňuje ve značné míře volba počtu tříd *L* a všech délek intervalů Δx_j. Pro přibližně symetrická rozdělení výběru lze

počítat L podle vztahu $L = \text{int}(2\sqrt{n})$, kde funkce int($x$) označuje celočíselnou část čísla x a v širokém rozmezí velikostí výběrů n užijeme výraz $L = \text{int}(2.46\ (n-1)^{0.4})$.

Jádrový odhad hustoty pravděpodobnosti (osa x: x, osa y: hustota pravděpodobnosti). Pro malé a střední výběry se konstruují jádrové odhady hustoty podle vztahu

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

kde šířka pásu h určuje stupeň vyhlazení. Jádrová funkce $K(x)$ je symetrická kolem nuly a má všechny vlastnosti hustoty pravděpodobnosti. O kvalitě odhadu hustoty pravděpodobnosti rozhoduje volba parametru h. Pro výběry velikosti n z přibližně normálního rozdělení se známým rozptylem σ^2 je optimální šířka pásu $h_{opt} = 2.34\ \sigma\ n^{-0.2}$.

Kvantil-kvantilový graf (graf Q-Q) (osa x: $Q_S(P_i)$, osa y: $x_{(i)}$). Umožňují posoudit shodu výběrového rozdělení, jež je charakterizováno kvantilovou funkcí $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$. Jako odhad kvantilové funkce výběru se využívají pořádkové statistiky $x_{(i)}$. Při shodě výběrového rozdělení se zvoleným teoretickým rozdělením platí přibližná rovnost kvantilů $x_{(i)} \approx Q_T(P_i)$, kde P_i je pořadová pravděpodobnost a závislost $x_{(i)}$ na $Q_T(P_i)$ je přibližně přímka. Korelační koeficient r_{xy} je pak kritériem těsnosti proložení této přímky při hledání typu neznámého rozdělení. Pro porovnání rozdělení výběru s rozdělením normálním se Q-Q graf nazývá *graf rankitový*.

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

Kruhový graf slouží k vizuálnímu ověření hypotézy, že výběr pochází ze symetrického (nejčastěji Gaussova) rozdělení. V takovém případě je grafem regulární, konvexní polygon, blízký kružnici. Odchylky od kružnice ukazují na jiné než symetrické rozdělení výběru: (a) protáhlý elipsovitý tvar s hlavní osou, umístěnou úhlopříčně ukazuje na asymetrické rozdělení, (b) elipsovitý tvar podél x-ové osy ukazuje na rovnoměrné rozdělení.

2.2 Ověření předpokladů o datech

V praxi se nejčastěji předpokládá *náhodný výběr* o velikosti n, tj. $\{x_i\}$, $i = 1, \ldots, n$. *Reprezentativní náhodný výběr* je charakterizován třemi důležitými předpoklady, které je třeba před vlastní analýzou ověřit. Jsou to nezávislost, homogenita a případná normalita výběru.

2.3 Transformace dat

Pokud se na základě analýzy dat zjistí, že rozdělení výběru dat se systematicky odlišuje od rozdělení normálního, vzniká problém, jak data vůbec vyhodnotit. Pak je často vhodná transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě. *Zesymetričtění rozdělení* výběru je možné provést užitím **jednoduché (prosté) mocninné transformace**

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

Mocninná transformace však nezachovává měřítko, není vzhledem k hodnotě λ všude spojitá a hodí se pouze pro kladná data. Optimální odhad exponentu λ se hledá s ohledem na optimalizaci

charakteristik asymetrie, (šikmosti) a špičatosti. K určení optimálního λ lze užít *selekčního grafu* dle Hinese a Hinesové.

Hinesové-Hinesův selekční graf (osa *x*: $\tilde{x}_{0.5}/x_{1-Pi}$, osa *y*: $\tilde{x}_{Pi}/\tilde{x}_{0.5}$). Vychází z požadavků symetrie jednotlivých kvantilů kolem mediánu $\left(\tilde{x}_{P_i}/\tilde{x}_{0.5}\right)^{\lambda} + \left(\tilde{x}_{0.5}/\tilde{x}_{1-P_i}\right)^{-\lambda} = 2$, kde jako kvantily jsou obvykle voleny písmenové hodnoty. Podle umístění experimentálních bodů v okolí teoretických křivek selekčního grafu lze odhadovat velikost λ a posuzovat kvalitu transformace v různých vzdálenostech od mediánu.

Pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti se užívá **Boxovy-Coxovy transformace**

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

Boxova-Coxova transformace je použitelná pouze pro kladná data. Rozšíření této transformace na oblast, kdy rozdělení dat začíná od prahové hodnoty x_0, spočívá v náhradě *x* rozdílem $(x - x_0)$, který je vždy kladný.

Graf logaritmu věrohodnostní funkce (osa *x*: λ, osa *y*: ln *L*). Pro odhad λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti s tím, že pro $\lambda = \hat{\lambda}$ je rozdělení transformované veličiny *y* normální, $N(\mu_y, \sigma^2(y))$. Po úpravách bude logaritmus věrohodnostní funkce ve tvaru

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

kde $s^2(y)$ je výběrový rozptyl transformovaných dat *y*. Průběh věrohodnostní funkce $\ln L = f(\lambda)$ lze znázornit ve zvoleném intervalu např. $-3 \leq \lambda \leq 3$ a identifikovat maximum křivky, jehož *x*-ová souřadnice indikuje $\hat{\lambda}$. Dva průsečíky křivky $\ln L(\lambda)$ s rovnoběžkou s x-ovou osou indikují 100(1-α)%ní interval spolehlivosti parametru λ. Čím bude tento interval spolehlivosti širší, tím je mocninná Boxova-Coxova transformace méně výhodná. Pokud obsahuje tento interval i hodnotu λ = 1, není transformace ze statistického hlediska přínosem.

Zpětná transformace: po vhodné transformaci určíme $\bar{y}$, $s^2(y)$ a potom pomocí zpětné transformace s využitím Taylorova rozvoje v okolí $\bar{y}$ odhadneme retransformované parametry $\bar{x}_R$ a $s^2(\bar{x}_R)$ původních dat. Uvedený postup vede vesměs k lepším odhadům polohy a rozptylu zvláště v případech asymetrického rozdělení.

3. Průběh průzkumové analýzy dat

Průběh vlastní průzkumové, exploratorní analýzy dat (EDA) je možné libovolně kombinovat. Předpokládá se, že rozdělení dat je normální a data asi splňují předpoklady nezávislosti a homogenity. Účelem je a) testování nezávislosti prvků výběru - autokorelace, b) testování homogenity výběru, c) testování normality rozdělení výběru. Z grafických metod se k předběžné analýze rutinních dat nejčastěji užívá *rankitových grafů* a *grafů rozptýlení s kvantily*. Nejsou-li však o rozdělení dat dostupné žádné informace nebo očekává-li se výrazně nenormální rozdělení, je vhodné provést a) průzkumovou analýzu dat s využitím grafických diagnostik, b) určení výběrového rozdělení a jeho konstrukci. Pokud nebylo nalezeno vhodné aproximující rozdělení, provádí se **vždy** *mocninná transformace* nebo *Boxova-Coxova transformace*, která by měla zlepšit rozdělení dat. Kombinace metod závisí na konkrétních datech a konkrétních požadavcích analýzy.

4. Ilustrativní příklad

Na příkladu *kontroly obsahu ergosterinu v calciferolu* ukážeme využití postupu průzkumové analýzy dat a vyčíslení střední hodnoty. Obsah ergosterinu [%]: 0.120, 0.188, 0.680, 0.140, 0.224, 0.000, 0.321, 0.050, 0.066, 0.670, 0.255, 0.318, 0.077, 0.150, 0.210, 0.032, 0.340, 0.150, 0.410, 0.074, 0.169, 0.301, 0.080, 0.100, 0.043, 0.320, 0.151, 0.094, 0.045, 0.140, 0.130, 0.130, 0.000, 0.160, 0.290, 0.058, 0.033, 0.200, 0.490, 0.000, 0.183, 0.069, 0.114.

Řešení:

1. Zkoumání zvláštností dat: grafické diagnostiky indikují vedle stupně symetrie a špičatosti rozdělení také odlehlé body.

(a) **Odhalení stupně symetrie a špičatosti rozdělení:** celkem 12 grafických diagnostik indikuje symetrii a špičatost rozdělení s těmito závěry:

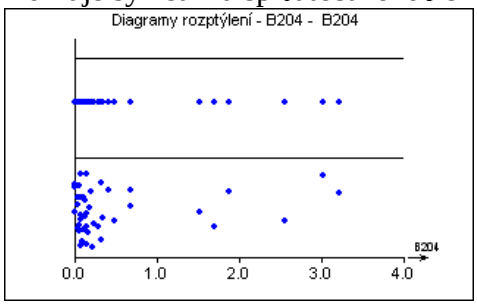
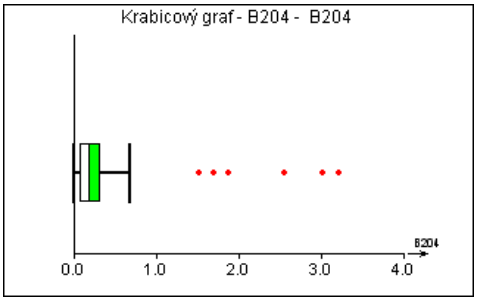
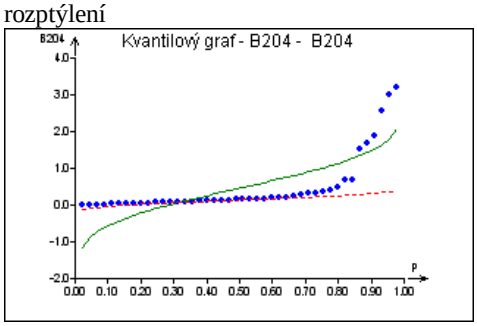


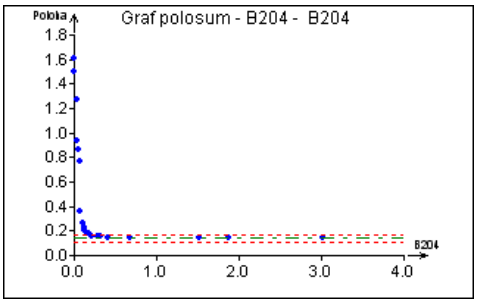
Diagram rozptýlení a rozmítnutý diagram rozptýlení



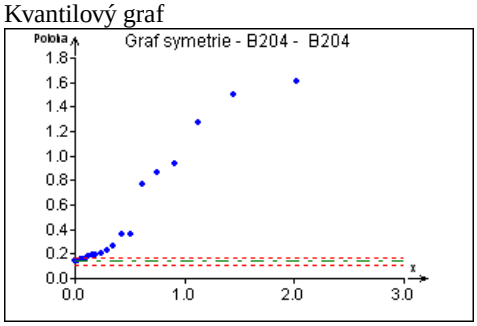
Krabicový graf



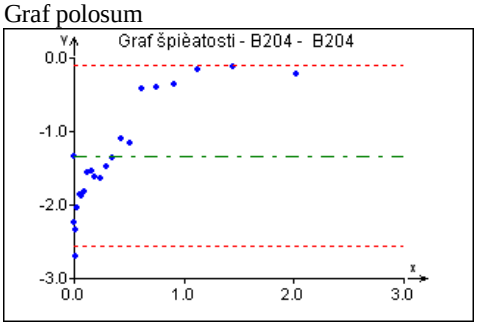
Kvantilový graf



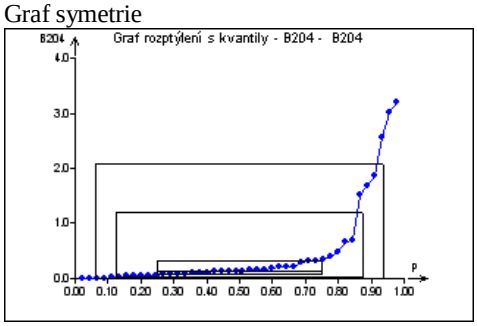
Graf polosum



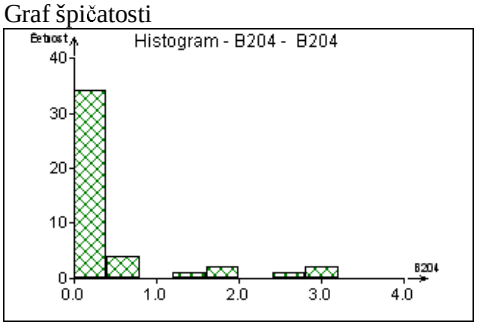
Graf symetrie



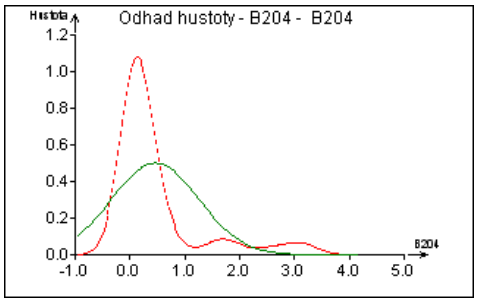
Graf špičatosti



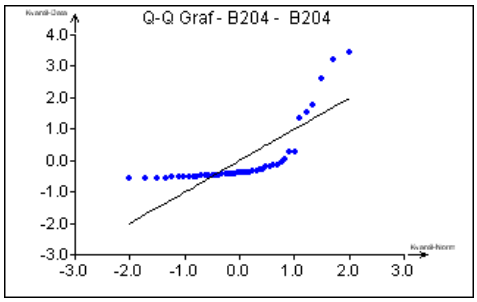
Graf rozptýlení s kvantily



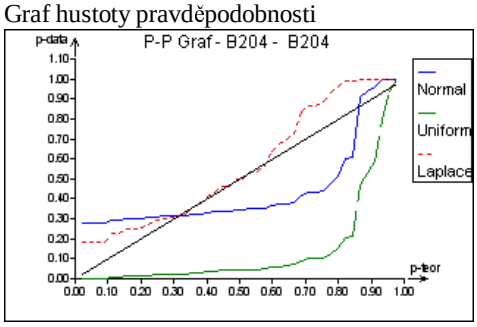
Histogram



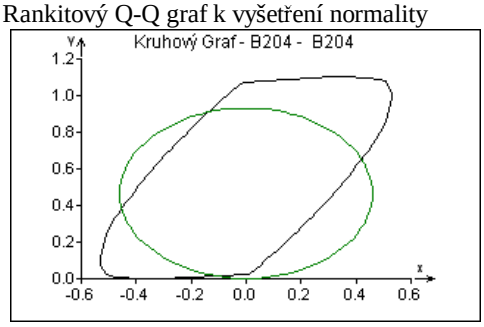
Graf hustoty pravděpodobnosti



Rankitový Q-Q graf k vyšetření normality



P-P graf



Kruhový graf

1. *Diagram rozptýlení a rozmítnutý diagram rozptýlení:* ukazuje na 4 až 6 odlehlých bodů v horní části diagramu.
2. *Krabicový (vrubový) graf:* v horní části je detekováno 6 odlehlých bodů. Krabice je rozdělena na dvě části mediánem. Je vyznačen interval spolehlivosti mediánu.
3. *Kvantilový graf:* je patrný velký rozdíl mezi symetrickým Gaussovým a empirickým rozdělením. Tvar křivky je charakteristický pro asymetrické rozdělení, silně sešikmené k vyšším hodnotám.
4. *Graf polosum:* indikuje značnou část bodů jako vybočujících ze symetrického rozdělení. Body ležící na mediánové rovnoběžce s x-ovou osou pocházejí ze symetrického rozdělení, ostatní body nikoliv.
5. *Graf symetrie:* indikuje valnou část bodů jako vybočujících ze symetrického rozdělení nebo body patřící do asymetrického rozdělení.
6. *Graf špičatosti:* většina bodů neleží na rovnoběžce s x-ovou osou pro symetrické rozdělení, a proto je indikováno rozdělení asymetrické.
7. *Graf rozptýlení s kvantily:* asymetrie kvantilových obdélníků obdélníků dokazuje silně asymetrické rozdělení. Body ležící vně sedecilového obdélníka indikuje tato pomůcka jako body odlehlé.
8. *Histogram:* zřetelně ukazuje na asymetrické rozdělení sešikmené k vyšším hodnotám.
9. *Jádrový odhad hustoty pravděpodobnosti:* ve srovnání s Gaussovým rozdělením je u empirické křivky patrné silné sešikmení k vyšším hodnotám. Empirickou křivku nelze aproximovat symetrickým Gaussovým rozdělením.
10. *Rankitový Q-Q graf:* jelikož většina bodů neleží na přímce Gaussova rozdělení jde o asymetrické rozdělení.
11. *Pravděpodobnostní P-P graf:* empirická křivka nesouhlasí s žádnou křivkou symetrického rozdělení (normálního, rovnoměrného a Laplaceova). Rozdělení je proto asymetrické.
12. *Kruhový graf:* tvar elipsy dokazuje silně asymetrické rozdělení sešikmené k vyšším hodnotám.

V exploratorní analýze dat umožňují kvantily a písmenové hodnoty posoudit jednak symetrii výběrového rozdělení a jednak procento prvků ve výběru, např. pro *procento* 45 je *kvantil* 0.1300, což znamená, že pod hodnotou 0.1300 leží 45% a nad 55% prvků výběru.

(b) **Indikace lokální koncentrace dat a rozdělení výběru:** aktuální výběrové rozdělení je asymetrické, s dlouhým horním koncem s větší koncentrací bodů ve spodní části hodnot. Z analýzy kvantil-kvantilového *Q-Q* grafu vyplývá, že nejvyšší hodnoty korelačního koeficientu r_{xy} = 0.98963 je dosaženo pro exponenciální rozdělení.

Určení výběrového rozdělení na základě *Q-Q* grafu (ADSTAT)

Linearita kvantil-kvantilovém (Q-Q) grafu $y = \beta_0 + \beta_1 x$:			
Rozdělení	**Směrnice β_0**	**Úsek β_1**	**Korelační koeficient r_{xy}**
Laplaceovo	1.1077E-01	1.7671E-01	9.2263E-01
Normální	1.4976E-01	1.7671E-01	9.2054E-01
Exponenciální	1.6708E-01	1.2594E-02	9.8963E-01
Rovnoměrné	4.8940E-01	-6.7991E-02	8.9212E-01
Lognormální	9.0554E-02	3.4126E-02	9.6956E-01

(c) **Nalezení vybočujících prvků ve výběru:** z grafických diagnostik průzkumové (exploratorní) analýzy výběru byly nalezeny 3 až 6 podezřelých bodů, které by mohly být chápany v symetrickém rozdělení jako odlehlé. Jelikož však jde o asymetrické rozdělení exponenciální, nemá smyslu indikovat odlehlé body.

2. Ověření předpokladů o datech: *Reprezentativní náhodný výběr* je charakterizován třemi důležitými předpoklady, které je třeba před vlastní analýzou ověřit. Z klasických odhadů byl vyčíslen odhad aritmetického průměru $\bar{x}$ = 0.177 a odhad směrodatné odchylky s = 0.159. Odhad šikmosti $\hat{g}_1$ = 1.54 a odhad špičatosti je $\hat{g}_2$ = 5.36 ukazoval na sešikmené rozdělení.

(a) **Ověření normality rozdělení:** Na předpokladu normality je založena celá statistická analýza dat. Test kombinace výběrové šikmosti a špičatosti ukázal , že normalita výběrového rozdělení je zamítnuta na vypočtené hladině významnosti 1.6254E-08.

(b) **Ověření nezávislosti prvků výběru:** K identifikaci časové závislosti prvků výběru nebo jiné vnitřní závislosti se testuje významnost autokorelačního koeficientu prvního řádu podle von Neumannova testovacího kritéria t_n = 0.4049, které bylo nižší než kritická hodnota $t_{1-\alpha/2}(n+1)$ = 2.0141, a proto nezávislost prvků ve výběru byla prokázána.

(c) **Ověření homogenity rozdělení výběru:** Homogenní výběr znamená, že všechny jeho prvky pocházejí ze stejného rozdělení s konstantním rozptylem. Vybočující měření silně zkreslují odhady polohy a zejména rozptylu s^2, takže zcela znehodnocují další statistickou analýzu. Testování vybočujících měření bez doplňkových informací průzkumové analýzy dat je málo spolehlivé. Kritérium vnitřních mezí určilo 2 odlehlé body, a to bod č. 14 a 23. Doplněním informací z průzkumové analýzy lze identifikovat exponenciální rozdělení výběru, které již odlehlé body mít nebude, takže toto vyloučení dvou bodů nemá statistický smysl.

(d) **Určení minimálního rozsahu výběru:** Uvažujeme-li 25% relativní chybu směrodatné odchylky, bude minimální rozsah výběru n = 18, pro 10% relativní chybu směrodatné odchylky pak n = 110 a pro 5% relativní chybu směrodatné odchylky bude n = 437.

3. Transformace dat: Jelikož se na základě průzkumové analýzy dat zjistilo, že rozdělení výběru dat se silně odlišuje od rozdělení normálního, vyvstává zde problém, jak vůbec vyhodnotit odhad střední hodnoty, když nelze použít aritmetický průměr. V takovém případě je vhodná transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a v případě Box-Coxovy transformace i k normalitě.

(a) **Mocninná transformace:** Zesymetričtění rozdělení výběru je možné provést užitím jednoduché (prosté) mocninné transformace. Pro odhad exponentu λ se hledají optimální hodnoty charakteristik asymetrie (šikmosti) a špičatosti. K určení optimálního λ lze ale také užít orientační grafické metody, selekčního grafu dle Hinese a Hinesové. Podle umístění experimentálních bodů v okolí teoretických křivek selekčního grafu byla odhadnuta velikost exponentu 0.5.

(b) **Box-Coxova transformace:** Pro přiblížení rozdělení normálnímu vzhledem k šikmosti a špičatosti se užívá Boxovy-Coxovy transformace. Pro odhad parametru λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti s tím, že pro $\lambda = \hat{\lambda}$ je rozdělení transformované veličiny *y* nejblíže normálnímu, $N(\mu_y, \sigma^2(y))$. Průběh věrohodnostní funkce $\ln L = f(\lambda)$ lze znázornit ve zvoleném intervalu např. $-3 \le \lambda \le 3$ a identifikovat maximum křivky v grafu logaritmu věrohodnostní funkce tak, že *x*-ová souřadnice indikuje nejlepší odhad $\hat{\lambda}$. Dva průsečíky křivky věrohodnostní funkce $\ln L(\lambda)$ s rovnoběžkou s x-ovou osou indikují $100(1-\alpha)$%ní interval spolehlivosti parametru λ, tj. $\langle \lambda_D, \lambda_H \rangle$. Jelikož tento interval spolehlivosti neobsahuje číslo +1, je mocninná a Boxova-Coxova transformace ze statistického hlediska výhodná a má smyl ji užívat. Po vhodné transformaci se určí $\bar{y}$, $s^2(y)$ a potom pomocí zpětné transformace s využitím Taylorova rozvoje v okolí $\bar{y} = 0.35465$ se odhadne retransformované parametry $\bar{x}_R = 0.14318$ a $s^2_R = 0.020931$ původních dat. Uvedený postup vede vesměs k lepším odhadům polohy a rozptylu a je vhodný zvláště v případech takového asymetrického rozdělení výběru.

Mocninná a Box-Coxova transformace (ADSTAT)

Prostá mocninná transformace, a **Box-Coxova transformace** (dává stejné výsledky):	
Odhad optimálního exponentu $\hat{\lambda}$	0.53
Opravený odhad průměru původních dat $\bar{x}_R$	0.143
Opravený odhad směrodatné odchylky původních dat s_R	0.145
Spodní mez intervalu spolehlivosti původních dat L_D	0.102
Horní mez mez intervalu spolehlivosti původních dat L_H	0.190

4. Závěr: Pro asymetrické rozdělení **nelze užít** za odhad střední hodnoty aritmetický průměr $\bar{x}$ = 0.177 se směrodatnou odchylkou *s* = 0.159. Protože výběr pochází z exponenciálního rozdělení, je nejlepším odhadem střední hodnoty *retransformovaný průměr* $\bar{x}_R = 0.143$ s *retransformovanou směrodatnou odchylkou* $s_R = 0.145$. Konečně 95%ní interval spolehlivosti retransformovaného průměru bude $L_D = 0.102$ a $L_H = 0.190$. Tomuto rigoróznímu odhadu střední hodnoty se nejvíce blíží jeho *robustní odhad*, a to *medián* $\tilde{x}_{0.5} = 0.140$.

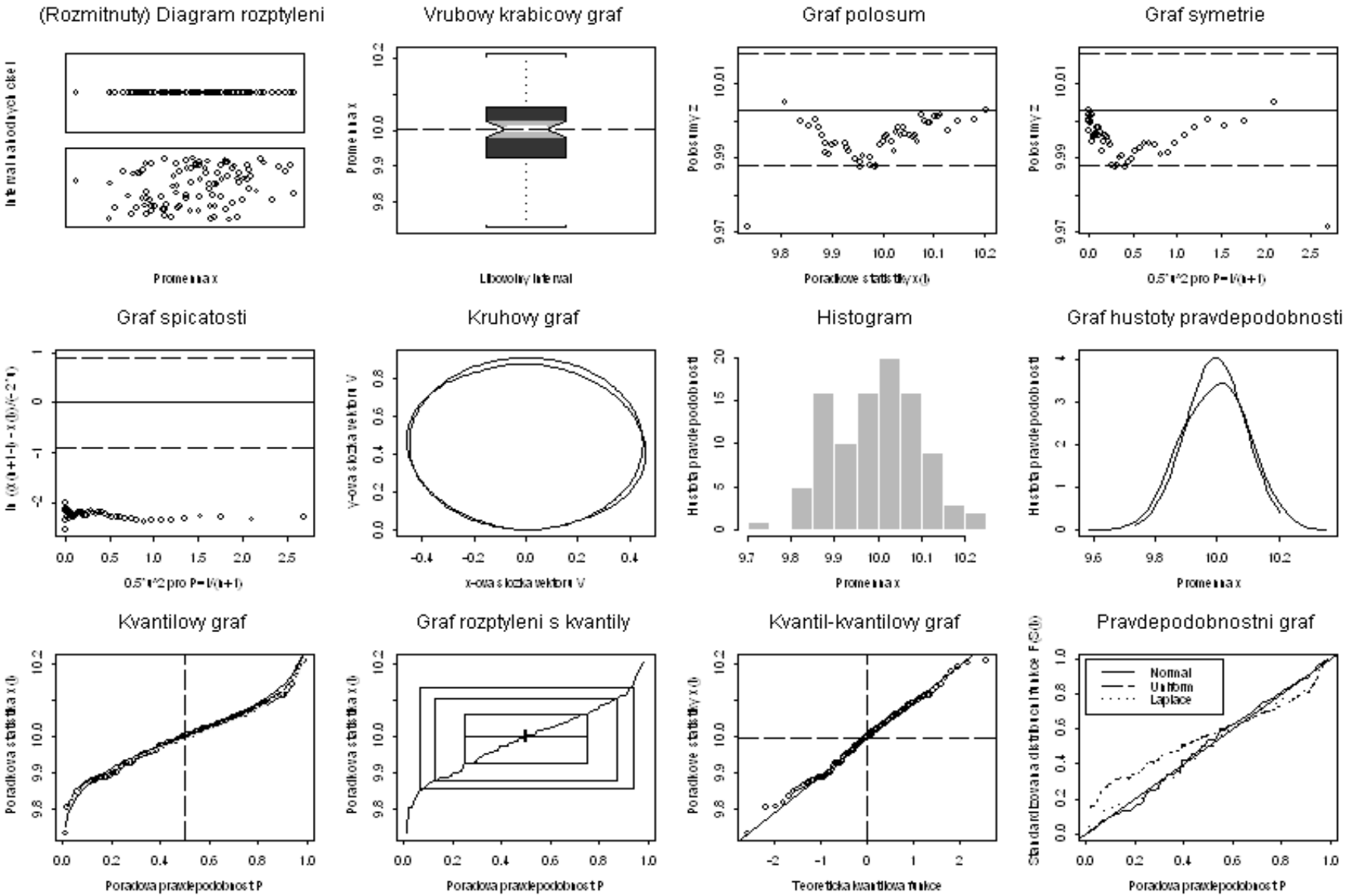
5. Doporučená literatura

[1] Meloun M., Militký J.: *Statistické zpracování experimentálních dat*, PLUS Praha 1994, ISBN 80-85297-56-6.

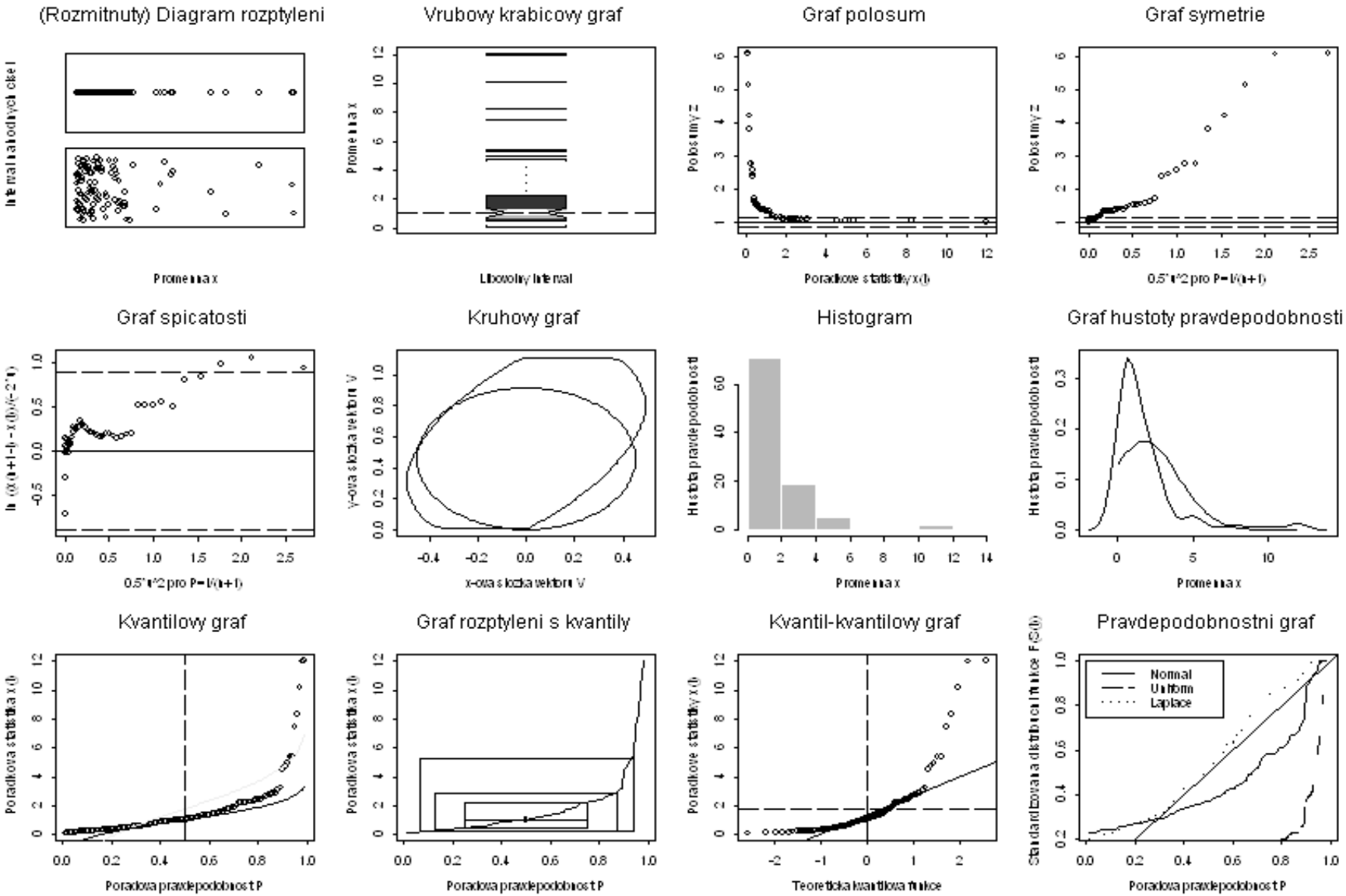
[2] Meloun M., Militký J.: *Statistické zpracování experimentálních dat - Sbírka úloh s disketou*, Univerzita Pardubice 1997, ISBN 80-7194-075-5.

[3] Kupka K.: *Statistické řízení jakosti*, Trilobyte Pardubice 1998, ISBN 80-238-1818-X.

[4] Militký J.: *Moderní statistické metody pro životní prostředí*, PHARE, Svazek 15, Vysoká škola báňská, Ostrava 1996, ISBN 80-7078-360-5.



Obr. 1 Přehled EDA diagnostik při analýze výběru normálního rozdělení N(10, 0.1), ***SPlus***



Obr. 2. Přehled EDA diagnostik při analýze výběru logaritmicko-normálního rozdělení LN(5, 2), ***SPlus***