2

PRŮZKUMOVÁ ANALÝZA JEDNOROZMĚRNÝCH DAT

Experimentální data se v analytické laboratoři často vyznačují nekonstantním rozptylem, malou četností, asymetrickým rozdělením a porušením základních předpokladů, kladených na výběr. Uveďme nejprve 3 etapy obecné osnovy analýzy výběru dat.

A. V průzkumové analýze dat se vyšetřují *statistické zvláštnosti dat*, jako je lokální koncentrace dat, tvarové zvláštnosti rozdělení dat a přítomnost podezřelých hodnot. Odhalí se také anomálie a odchylky rozdělení výběru od typického rozdělení, obyčejně normálního či Gaussova. Interaktivní statistická analýza na počítači tento postup ulehčuje, většina statistického software nabízí řadu diagnostických grafů a diagramů. Pokud je rozdělení dat nevhodné pro standardní statistickou analýzu (tj. většinou je asymetrické), provádí se nejprve vhodná transformační úprava dat. Pokud bylo indikováno sešikmené rozdělení nebo rozdělení s dlouhými konci, pomocníkem je mocninná a Boxova-Coxova transformace. Transformace je vhodná především při asymetrii rozdělení původních dat, ale také při nekonstantnosti rozptylu.

B. Pro případ rutinních měření 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. obyčejně aritmetického průměru a rozptylu. Dále se vyčíslí intervaly spolehlivosti, následované testováním statistických hypotéz. V pesimistickém případě následuje další pokus o úpravu dat.

C. V konfirmatorní analýze je nabízena paleta rozličných odhadů polohy, rozptýlení a tvaru, jež lze rozdělit do dvou skupin: na *klasické odhady* a na *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 ověření předpokladů o výběru.

Postup statistické analýzy jednorozměrných dat[1, 3, 5], prováděné v interaktivním režimu na počítači lze shrnout do bloků operací, i když lze jednotlivé operace provádět samostatně: operace A, operace B, operace A+B, operace B+C a konečně všechny operace A+B+C.

Přehled operací analýzy jednorozměrných dat

A. Průzkumová (exploratorní) analýza dat (EDA):

- Odhalení stupně symetrie a špičatosti výběrového rozdělení;
- Indikace lokální koncentrace výběru dat;
- Nalezení vybočujících a podezřelých prvků ve výběru;
- Porovnání výběrového rozdělení dat s typickými rozděleními;
- Mocninná transformace výběru dat;
- Box-Coxova transformace výběru dat.

B. Ověření předpokladů o datech:

- Ověření nezávislosti prvků výběru dat;
- Ověření homogenity rozdělení výběru dat;
- Určení minimálního rozsahu výběru dat;
- Ověření normality rozdělení výběru dat.

C. Konfirmatorní analýza dat (CDA) - odhady parametrů (polohy, rozptýlení a tvaru)

1. Klasické odhady (bodové a intervalové) výběru dat;
2. Robustní odhady (bodové a intervalové) výběru dat.

Vzorová úloha 2.1 *Analýza dat normálního a log.-normálního rozdělení*

Analýza simulovaných dat výběru, pocházejícího z (a) rozdělení normálního ***norm*** *N(10, 0.1)* a z (b) rozdělení logaritmicko-normálního ***log*** *L(5, 2)*.

Data:

(a) Výběr ***norm****:*

10.0010	9.9290	10.0370	9.9490	10.1850	9.9590	10.0630	9.8790	10.0500	9.8460
...	...	...	...	...	...	...	...	...	...
10.0370	10.0110	9.9310	9.9870	9.9550	10.0130	10.0020	10.1150	10.0250	

(b) Výběr ***log***:

2.408	5.389	2.259	2.439	2.173	1.157	0.892	0.498	0.351	1.229
...	...	...	...	...	...	...	...	...	...
2.816	0.666	4.972	0.451	1.316	3.241	0.316	2.200	8.291	0.815

Řešení: u jednotlivých diagnostických diagramů a grafů budou uvedeny vždy dvě ukázky, jednak pro výběr ze symetrického normálního rozdělení *norm N*(10, 0.1) a jednak pro výběr z asymetrického logaritmicko-normálního rozdělení *log L*(5, 2). Čtenář může porovnat, jak jednotlivé diagnostické pomůcky monitorují symetrické a silně asymetrické rozdělení.

2.1 Průzkumová (exploratorní) analýza dat EDA

Prvním krokem v analýze jednorozměrných dat je průzkumová čili exploratorní 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 $E(x_{(i)})$ je rovna $100P_i$ procentnímu kvantilu výběrového rozdělení $Q(P_i)$ a symbol $P_i \approx i/(n+1)$ označuje *pořadovou pravděpodobnost*. Připomeňme, že *$100P_i$ procentní výběrový kvantil* je hodnota, pod kterou leží $100P_i$ procent prvků výběru. Vynesením hodnot $x_{(i)}$ proti P_i, $i = 1, \ldots, n$, se získá hrubý odhad *kvantilové funkce* $Q(P_i)$. Ta je inverzní k *funkci distribuční* $F(x_i)$ a charakterizuje jednoznačně rozdělení výběru.

Pro libovolnou hodnotu α z intervalu [0, 1] lze vyčíslit 100α%ní kvantil $\tilde{x}_\alpha$ pomocí lineární interpolace

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

kde pro index *i* musí být splněna nerovnost $\frac{i}{n+1} \le \alpha \le \frac{i+1}{n+1}$. Pro rozptyl kvantilu $\tilde{x}_\alpha$, určeného z výběru velikosti *n*, platí vztah $D(\tilde{x}_\alpha) = \frac{\alpha(1-\alpha)}{n\,[f(\tilde{x}_\alpha)]^2}$, kde $f(\tilde{x}_\alpha)$ je výběrová hodnota hustoty pravděpodobnosti v bodě $\tilde{x}_\alpha$.

V průzkumové analýze se často používá speciálních *kvantilů L* pro pořadové pravděpodobnosti $P_i = 2^{-i}$, $i = 1, 2, \ldots$, které se také nazývají *písmenové hodnoty*.

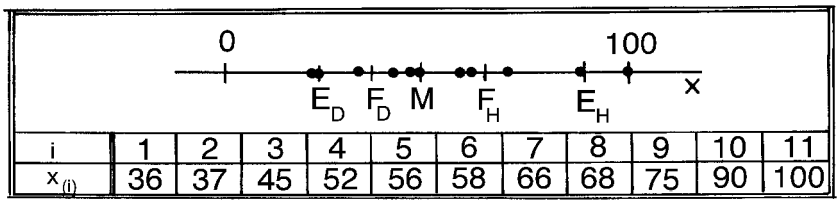
i	i-tý kvantil	Pořadová pravděpodobnost P_i	Písmenová hodnota *L*
1	Medián	$2^{-1} = 1/2$	*M*
2	Kvartily	$2^{-2} = 1/4$	*F*
3	Oktily	$2^{-3} = 1/8$	*E*
4	Sedecily	$2^{-4} = 1/16$	*D*

Symbol u_{P_i} označuje kvantil normovaného normálního rozdělení *N*(0, 1). Kromě *mediánu* (*i* = 1) existují pro každé *i* > 1 dvojice kvantilů, a to *dolní* a *horní písmenová hodnota* L_D a L_H. Dolní písmenová hodnota je pro pořadovou pravděpodobnost $P_i = 2^{-i}$, zatímco horní je pro $P_i = 1 - 2^{-i}$.

Pro odhad písmenových hodnot lze použít jednoduché techniky *pořadí* a *hloubek*. Pořádková statistika $x_{(i)}$ má *rostoucí pořadí* $R_{P_i} = i$ a *klesající pořadí* $K_{P_i} = n + 1 - i$. *Hloubka* H_i je pak menší číslo z obou pořadí $H_i = \min(R_{P_i}, K_{P_i})$. Pro hloubku mediánu platí $H_M = \frac{n+1}{2}$. Pokud je tato hloubka celé číslo, je medián $\tilde{x}_{0.5} = M = x_{(H_M)}$. V opačném případě se provádí lineární interpolace mezi $x_{(n/2)}$ a $x_{(n/2+1)}$. Hloubky dolních písmenných hodnot jsou $H_L = [1 + int(H_{L-1})]/2$, kde L jsou indexy F, E, D a int(x) značí celočíselnou část čísla x. Pokud je $L = F$, bere se $L - 1 = M$. Jestliže je H_L celé číslo, bude dolní kvantil $L_D = x_{(H_L)}$ a horní kvantil $L_H = x_{(n+1-H_L)}$. Je-li H_L číslo necelé, provádí se lineární interpolace podle vztahů

$$L_D = \frac{x_{int(H_L)} + x_{int(H_L)+1}}{2} \quad \text{a} \quad L_H = \frac{x_{n+1-int(H_L)} + x_{n+2-int(H_L)}}{2}.$$

Tento postup se pro menší hodnoty H_L, kdy jsou kvantily blízko hodnot $x_{(1)}$ a $x_{(n)}$, považuje za robustnější. Počet písmenových hodnot závisí na rozsahu výběru. Pro velikost výběru n lze určit n_L písmenových hodnot včetně mediánu. Platí, že $n_L \approx 1.44 \ln(n + 1)$.



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

(a)

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

(b) n=11

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

Obr. 2.1 Graficko-tabelární schéma sumarizace dat: (a) obecné schéma písmenově-číslicového zápisu výběru, (b) sedmipísmenový zápis výběru {36, 37, 45, 52, 56, 58, 66, 68, 75, 90, 100}.

Mezi základní statistické zvláštnosti rozdělení dat patří symetrie výběrového rozdělení a jeho relativní délky konců ve srovnání s normálním rozdělením. K vyjádření symetrie a špičatosti v různých vzdálenostech od mediánu se užívají jednoduché funkční charakteristiky, založené na písmenových hodnotách. Ze vztahu pro délku konců T_L v následující tabulce lze snadno určit jejich teoretické velikosti pro vybraná symetrická rozdělení, a to (T_E, T_D) hodnoty: *normální rozdělení* (0.534; 0.822), *rovnoměrné rozdělení* (0.405; 0.559) a *Laplaceovo rozdělení* (0.693; 1.098).

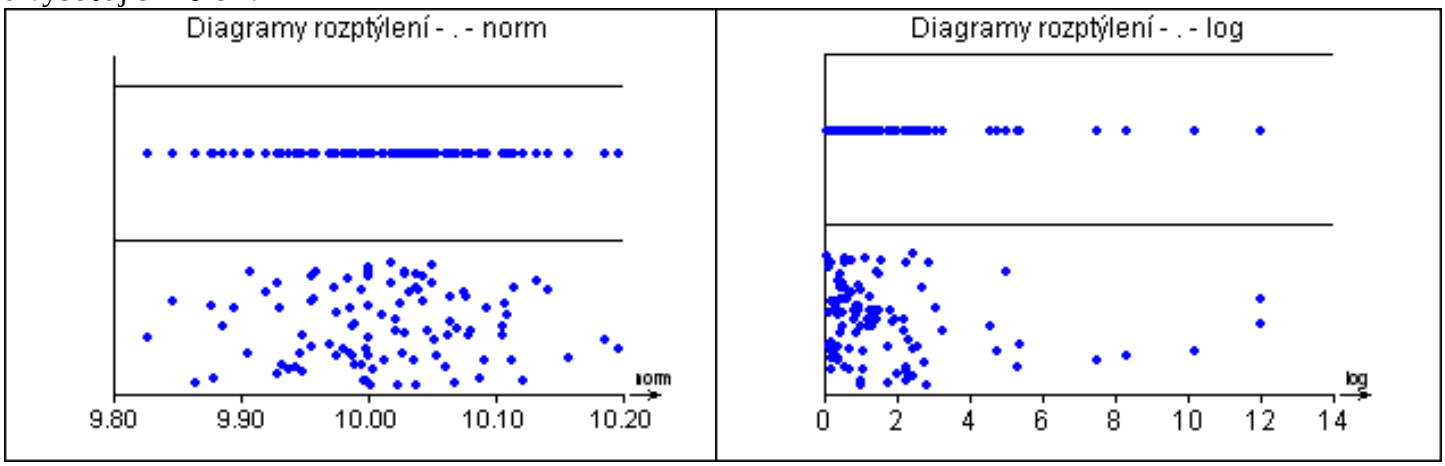
Pro rozdělení sešikmená k vyšším hodnotám jsou hodnoty šikmosti S_L záporné, při sešikmení k nižším hodnotám jsou kladné (viz tabulka Charakteristik šikmosti a špičatosti výběrového rozdělení). Pro rozdělení s delšími konci než má normální rozdělení rostou hodnoty pseudosigmy G_L se vzdáleností od mediánu. Když hodnoty G_L klesají s rostoucí vzdáleností od mediánu, má výběrové rozdělení kratší konce než normální. K posouzení statistických zvláštností dat se používá různých grafů, využívajících charakteristik z tabulky. Pro větší výběry se v grafech znázorňují pouze funkce písmenových hodnot, zatímco pro menší výběry se využívá všech kvantilů $\tilde{x}_{P_i} = x_{(i)}$ obyčejně při volbě $P_i = [i - 0.333]/[n + 0.333]$.

Charakteristiky šikmosti a špičatosti výběrového rozdělení

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

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

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.

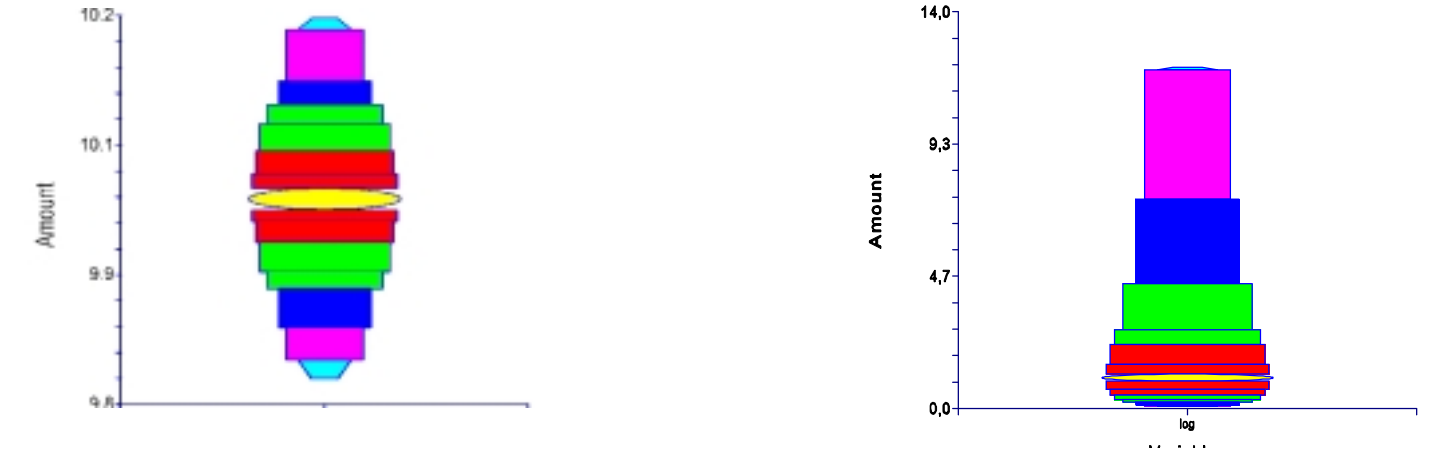


Obr. 2.2 Diagram rozptýlení a rozmítnutý diagram rozptýlení pro výběry (shora dolu): (a) *norm*, symetrického (Gaussova, normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT***

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

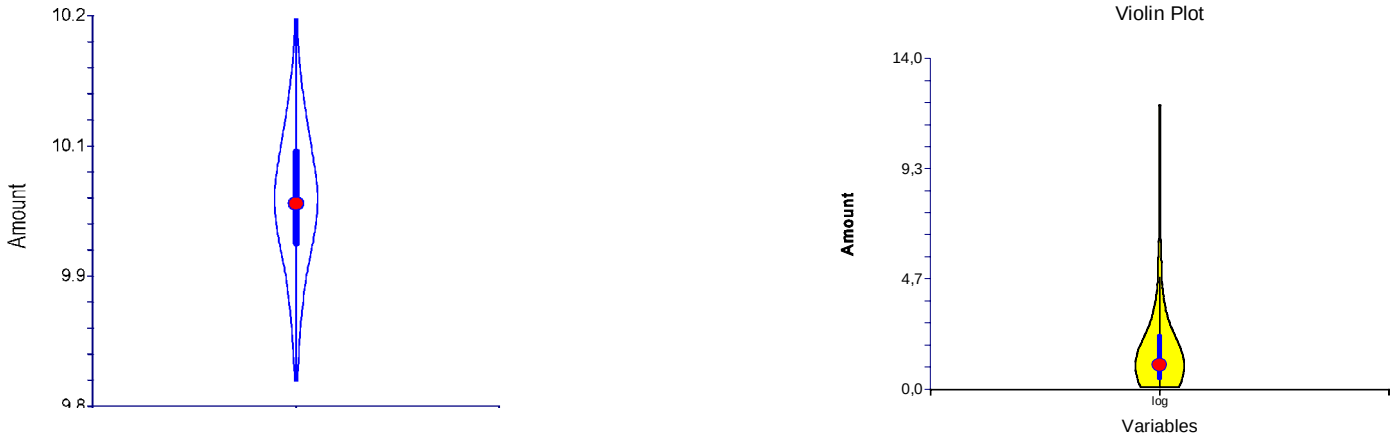
Obr. 2.3 Rozmítnutý diagram rozptýlení pro výběry: (a) *norm*, symetrického (Gaussova, normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000***

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



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

Houslový diagram (osa *x*: název výběru proměnné, osa *y*: percentily, hodnoty proměnné). Diagram je kombinací krabicového grafu a dvou vertikálních, zrcadlově k sobě zobrazených grafů hustoty. Jeden graf hustoty roste směrem doprava a druhý doleva. Diagram zobrazuje píky a údolí stejně jako graf hustoty pravděpodobnosti. Medián je zobrazen černým kolečkem a začátek a konec úsečky zobrazuje dolní a horní kvantil. Houslový diagram se jmenuje dle připomínajícího tvaru houslí. Normální rozdělení se projeví v symetrickém tvaru houslí zatímco log.-normální v silně asymetrickém tvaru.



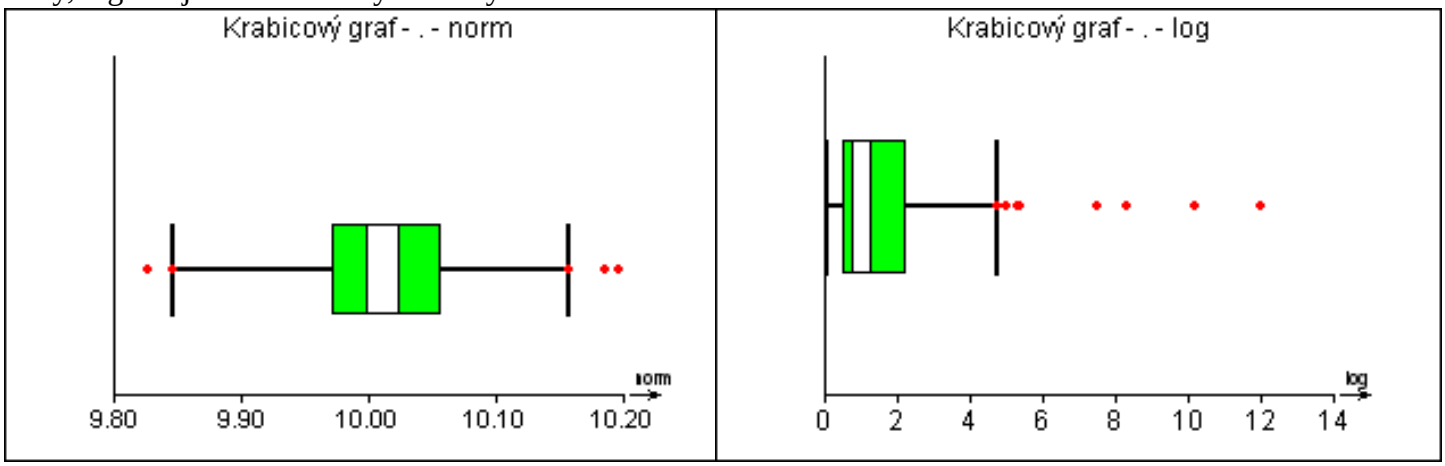
Obr. 2.5 Houslový diagram pro výběry: (a) *norm*, symetrického (Gaussova, normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000***

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

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

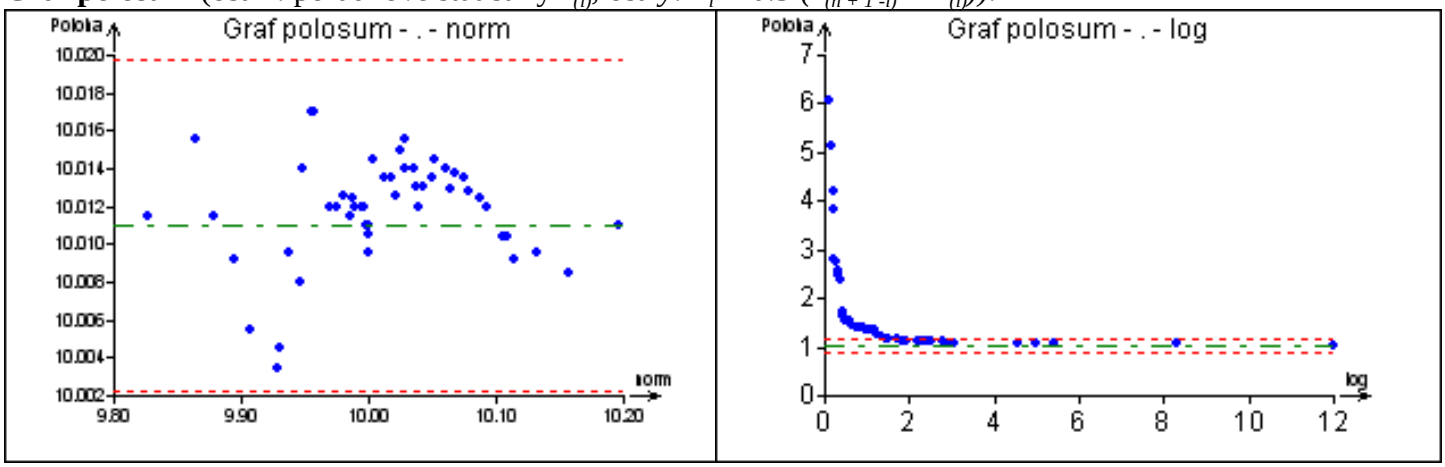
Prvky výběru, ležící mimo interval vnitřních hradeb [B_H, B_D,] jsou považovány za podezřelé, obvykle vybočující body, v grafu jsou znázorněny kroužky.



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

Vrubový krabicový graf (osa *x*: úměrná hodnotám x_i, osa *y*: interval úměrný hodnotě $\sqrt{n}$). Obdobou krabicového grafu je vrubový krabicový graf, který umožňuje také posouzení variability mediánu. Ta je totiž vyjádřena dolní a horní mezí intervalu spolehlivosti mediánu, $I_D \le M \le I_H$, který bývá znázorněn v okolí mediánu bílým proužkem.

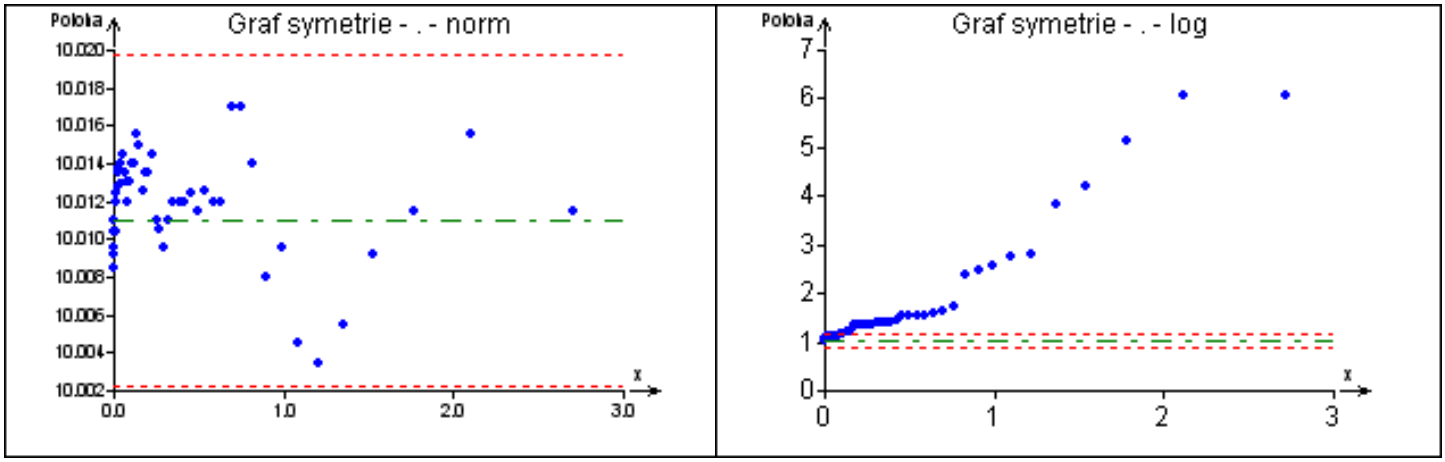
Graf polosum (osa *x*: pořádkové statistiky $x_{(i)}$, osa *y*: $Z_i = 0.5\ (x_{(n+1-i)} + x_{(i)})$).



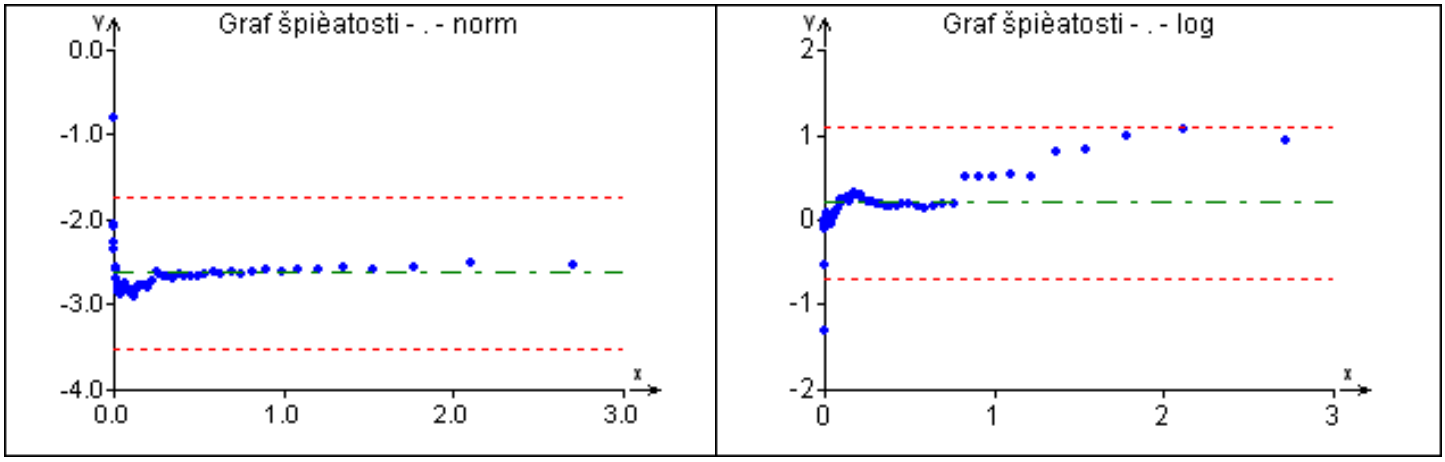
Obr. 2.8 Graf polosum pro výběry: (a) *norm*, symetrického (Gaussova, normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT***

Pro symetrické rozdělení je grafem polosum horizontální přímka, určená rovnicí $y = M$. U tohoto grafu je důležité, že zde body oscilují okolo horizontální přímky a vykazují tak náhodný shluk (mrak) a měřítko *y*-nové osy je silně detailní. Naopak asymetrické rozdělení vykazuje nenáhodný trend a body pak neosciluj okolo horizontální přímky a měřítko *y*-nové osy není detailní.

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



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

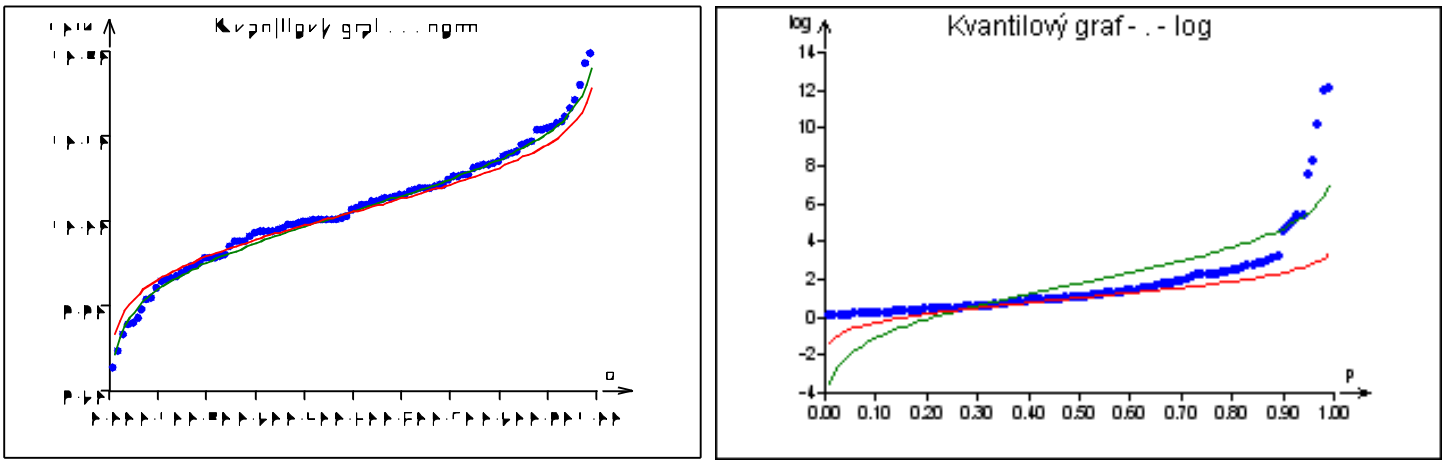


Obr. 2.11 Graf špičatosti pro výběry: (a) *norm*, symetrického (Gaussova, normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT***

Kvantilový graf (osa *x*: pořadová pravděpodobnost P_i, osa *y*: pořádková statistika $x_{(i)}$). Umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, které může být symetrické, sešikmené k vyšším nebo nižším hodnotám. Ke snadnějšímu porovnání s normálním rozdělením se do tohoto grafu zakreslují i kvantilové funkce normálního rozdělení, $N_{P_i} = \hat{\mu} + \hat{\sigma}\, u_{P_i}$, pro $0 \le P_i \le 1$:

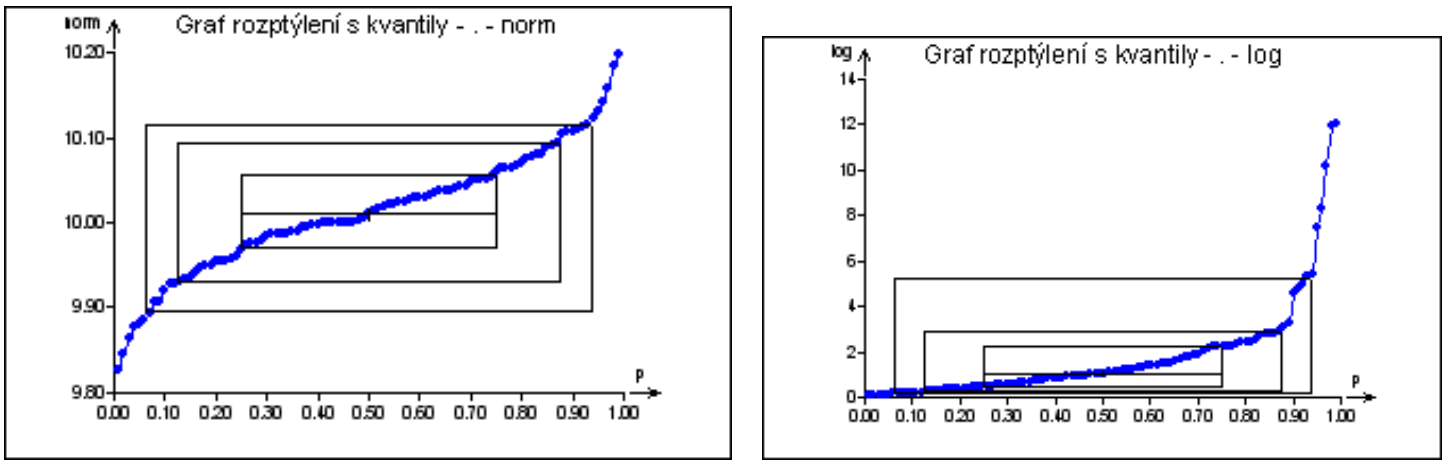
(1) *klasických odhadů* parametrů polohy a rozptýlení, tj. aritmetického průměru a směrodatné odchylky $\hat{\mu} = \bar{x}$ a $\hat{\sigma} = s$, a dále

(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.



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

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.



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

Do grafu se zakreslují tři pomocné kvantilové obdélníky:

(a) *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 $1 - 2^{-2} = 0.75$.

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

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

Graf rozptýlení s kvantily poskytuje následující závěry vyšetření dat:

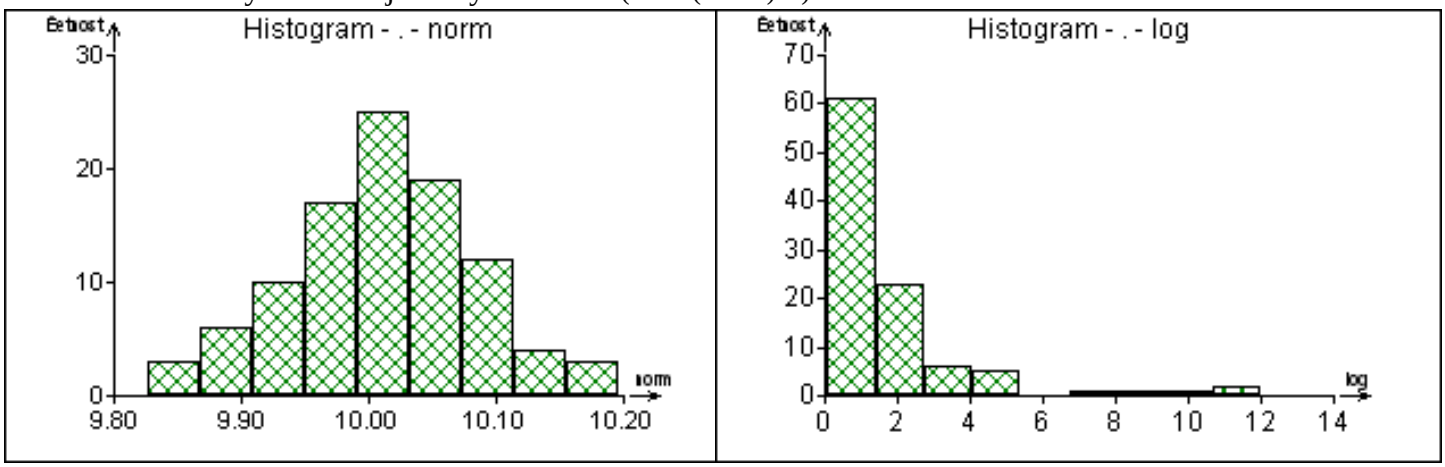
1. *Symetrické unimodální rozdělení výběru* obsahuje obdélníky symetricky uvnitř sebe.

2. *Nesymetrická rozdělení* mají pro rozdělení sešikmené k vyšším hodnotám vzdále-nosti mezi dolními hranami obdélníků *F*, *E* a *D* výrazně kratší než mezi jejich horními.

3. *Odlehlá pozorování* jsou indikována tím, že na kvantilové funkci mimo obdélník *D* se objeví náhlý vzrůst, kdy hodnota směrnice roste prakticky 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*. Pro přibližně symetrická rozdělení výběru lze vyčíslit počet tříd *L* podle vztahu $L = \text{int}(2\sqrt{n})$, kde funkce int(*x*) označuje celočíselnou část čísla *x*. V širokém rozmezí velikostí výběrů *n* užijeme výraz $L = \text{int}(2.46\ (n-1)^{0.4})$.



Obr. 2.13 Histogram pro výběry: (a) *norm*, symetrického (Gaussova, normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT***

Pro malé a střední výběry se konstruují jádrové odhady hustoty podle vztahu

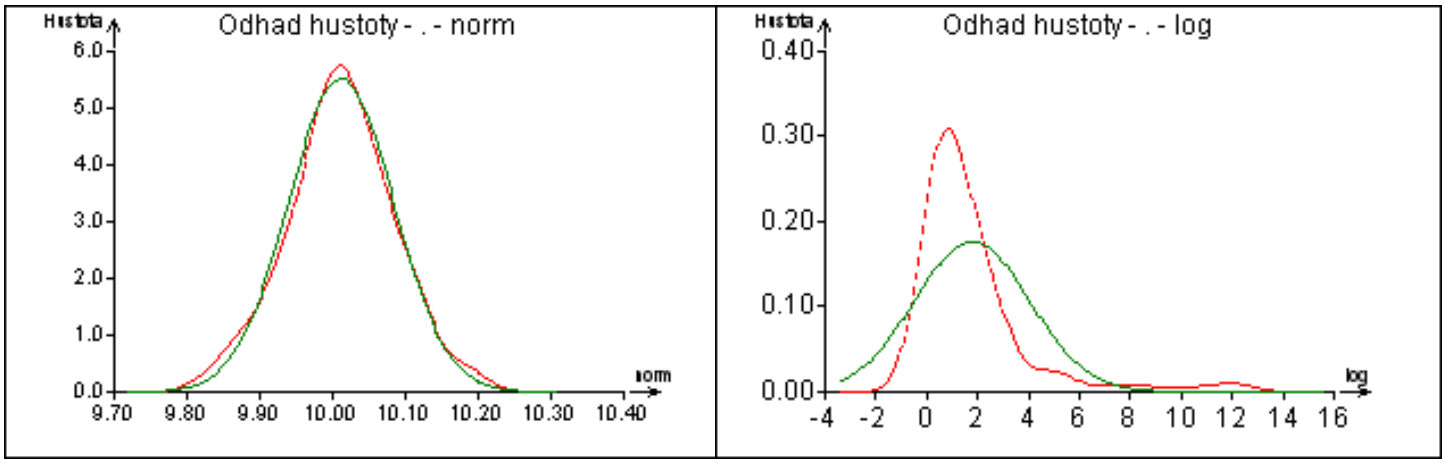
$$\hat{f}(x) = \frac{1}{n\,h}\sum_{i=1}^{n} K\left[\frac{(x - x_j)}{h}\right],$$ kde šířka pásu *h* určuje stupeň vyhlazení. Jádrová funkce *K(x)* je symetrická kolem nuly a má všechny vlastnosti hustoty pravděpodobnosti. Vhodná je t.zv. *bikvadratická funkce*

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

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

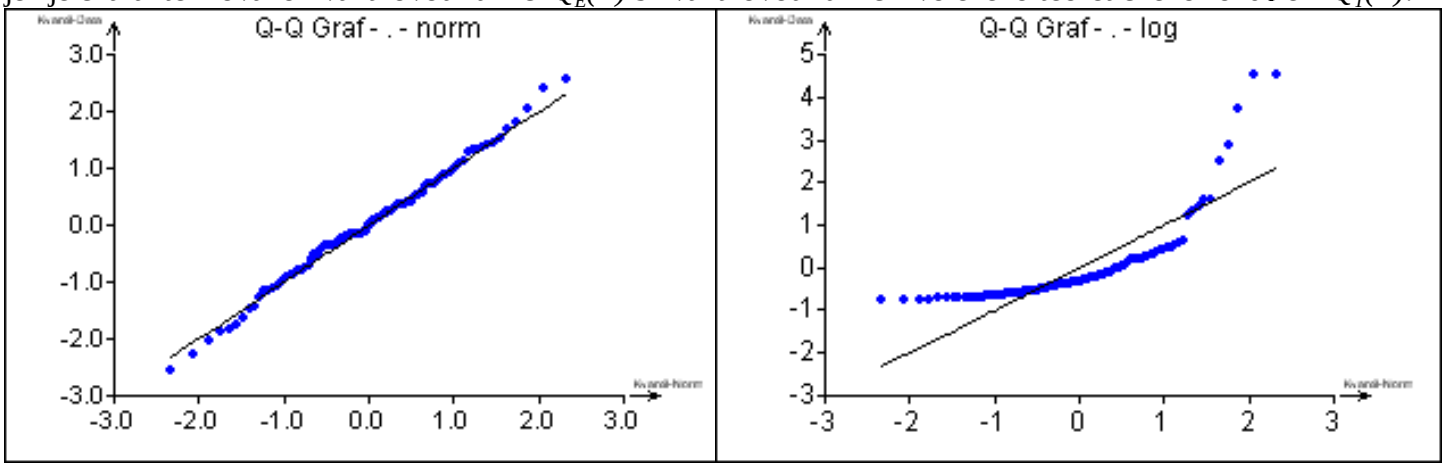
O kvalitě odhadu hustoty pravděpodobnosti rozhoduje volba parametru *h*. Pro výběry velikosti *n* z přibližně normálního rozdělení se známým rozptylem σ^2 je optimální šířka pásu $h_{opt} = 2.34\ \sigma\ n^{-0.2}$.

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



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

Kvantil-kvantilový graf (graf *Q*-*Q*) (osa *x*: $Q_T(P_i)$, osa *y*: $x_{(i)}$). Umožňují posoudit shodu výběrového rozdělení, jež je charakterizováno kvantilovou funkcí $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$.



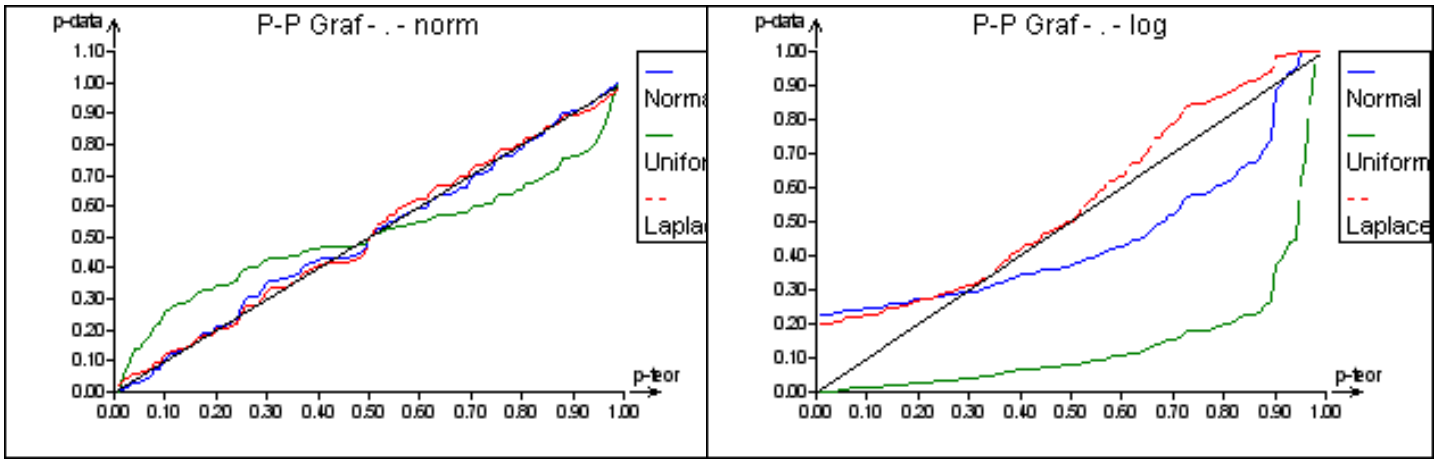
Obr. 2.15 Rankitový čili kvantil-kvantilový graf (Q-Q graf) pro ověření shody s teoretickým normálním rozdělením: (a) *norm*, symetrického (Gaussova, normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT***

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.

Rankitový graf (osa *x*: kvantil normovaného Gaussova rozdělení u_{P_i}, osa *y*: $x_{(i)}$). Pro porovnání rozdělení výběru s rozdělením normálním se *Q*-*Q* graf nazývá *grafem rankitovým*. Umožňuje také orientační zařazení výběrového rozdělení do skupin podle šikmosti, špičatosti a délky konců.

Podmíněný rankitový graf (osa *x*: $\Phi^{-1}\,[0.5\,(U_{(i-1)} + U_{(i+1)})]$, osa *y*: $x_{(i)}$). K ověření normality výběrového rozdělení se užívá podmíněný rankitový graf. Symbol $\Phi^{-1}(U)$ značí standardizovanou kvantilovou funkci normálního rozdělení. Hodnoty $U = P_i$ jsou přímo kvantily u_{P_i}. Pořádkové statistiky $U_{(i)}$ jsou uspořádané náhodné proměnné U_i definované vztahem $U_i = \Phi\left[x_i - \hat{\mu}_R / \hat{\sigma}_R\right]$, kde symbol $\Phi(x)$ značí distribuční funkci standardizovaného normálního rozdělení. Robustní odhad polohy je roven mediánu $\hat{\mu}_R = \tilde{x}_{0.5}$ a robustní směrodatná odchylka se vyčíslí vztahem $\hat{\sigma}_R = 0.75(\tilde{x}_{0.75} - \tilde{x}_{0.25})$. Pro úplnou definici se volí $U_{(0)} = 0$ a $U_{(n+1)} = 1$. 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.

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* a *R* je *parametr rozptýlení*. V případě shody výběrového rozdělení se zvoleným teoretickým rozdělením vyjde *P*-*P* graf lineární s jednotkovou směrnicí a nulovým úsekem.



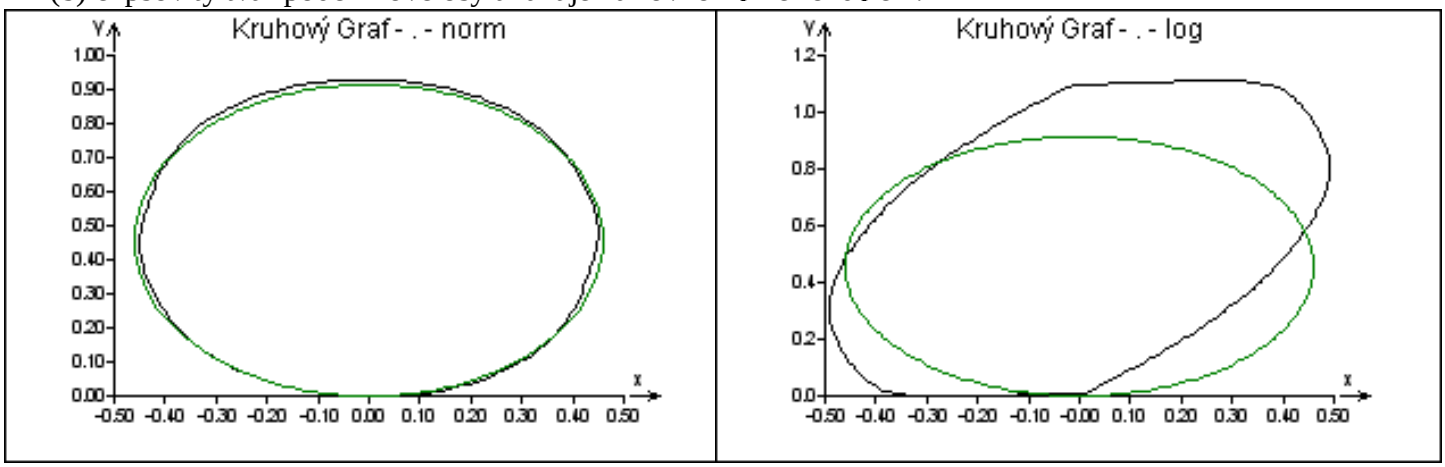
Obr. 2.17 Pravděpodobnostní graf (P-P graf) pro výběry: (a) *norm*, symetrického (Gaussova, normálního) a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT***

Na rozdíl od *Q-Q* grafů je při konstrukci *P-P* grafů nezbytné znát teoretické rozdělení až do hodnot všech parametrů. Obyčejně se určují odhady parametrů *Q* a *R* a navíc i dalších parametrů rozdělení s využitím momentové, resp. metody maximální věrohodnosti. Např. pro normální rozdělení je $\hat{R} = s$, $\hat{Q} = \bar{x}$. Při porovnání *Q-Q* a *P-P* grafů platí, že

a) *P-P grafy* jsou citlivé na odchylky od teoretického rozdělení ve střední části,
b) *Q-Q grafy* jsou citlivé na odchylky od teoretického rozdělení v oblasti konců.

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.



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

Vychází se z faktu, že transformací $Z_{(i)} = F_e(x_{(i)})$ vyjdou náhodné veličiny $Z_{(i)}$ rozdělené přibližně rovnoměrně na intervalu [0, 1]. Při konstrukci kruhového grafu se definuje soustava vektorů $\bar{V}_i$ o stejné délce $l_0 = 1/\sqrt{N\ (N - 1)/2}$ a směru $\pi\ Z_{(i)}$. Pro *x*-ovou a *y*-ovou složku vektoru $\bar{V}_i$ platí $V_{x_i} = l_0 \cos(\pi\ Z_{(i)})$, $V_{y_i} = l_0 \sin(\pi\ Z_{(i)})$. Úhly se uvažují v radiánech. Vlastní kruhový graf pak vznikne, když se postupně (od počátku) spojují vektory $\bar{V}_1$, $\bar{V}_2$, ... $\bar{V}_N$, - $\bar{V}_1$, - $\bar{V}_2$, ..., - $\bar{V}_N$. Výsledný obrazec je 2*N* vrcholový konvexní polygon. Odchylky od ideálního tvaru ukazují na nevhodnost specifikace F_e, resp. jeho parametrů. Obyčejně se jako F_e volí distribuční funkce normálního rozdělení $\Phi(x_{(i)})$ a kruhový graf pak slouží pro ověřování normality.

(1) Linearita kvantil-kvantilového (Q-Q) grafu $y = \beta_0 + \beta_1 x$:

Rozdělení	**Směrnice β_1**	**Úsek β_0**	**Korelační koeficient r_{xy}**
Laplaceovo	0.05246	10.012	0.99039
Normální	0.07289	10.012	0.99707
Exponenciální	0.06796	9.944	0.90972
Rovnoměrné	0.24233	9.891	0.97036
Lognormální	0.03253	9.960	0.82901
Gumbelovo	0.05623	10.044	0.97527

(2) Kvantilové míry polohy a rozptýlení:

Kvantil L	**P**	**Spodní L_D**	**Horní L_H**	**Rozpětí R_L**
Medián M	0.5	10.011	-	-
Kvartil F	0.25	9.9690	10.060	0.09100
Oktil E	0.125	9.9300	10.105	0.17450
Sedecil D	0.0625	9.8872	10.120	0.23225

(3) Vybrané míry polohy, rozptýlení a tvaru rozdělení, vyčíslené z kvantilů:

Kvantil L	**P**	**Polosuma Z_L**	**Šikmost S_L**	**Délka konců T_L**	**Pseudos.**
Kvartil F	0.25	10.015	-0.038462	0.00000	0.000000
Oktil E	0.125	10.017	-0.035817	-19.1851	0.51708
Sedecil D	0.0625	10.003	0.032831	-16.0430	0.80089

Korelační koeficient r_{xy} dosahuje nejvyšší hodnoty pro normální rozdělení, a tím dokazuje, že výběr *norm* pochází ze souboru s normálním rozdělením.

2.2 Ověření předpokladů o datech

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

1. předpoklad: **Prvky výběru x_i jsou vzájemně nezávislé**
Pokud se podmínky pro měření dat mění s časem, projeví se tyto vznikem trendu mezi prvky výběru, uspořádanými v časovém sledu. K identifikaci časové závislosti prvků výběru nebo závislosti související s pořadím jednotlivých měření, testuje se významnost autokorelačního koeficientu prvního řádu ρ_a podle testovacího von Neumannova kritéria

$$t_n = \frac{T_1 \sqrt{n+1}}{\sqrt{1 - T_1}} \text{ , kde } T_1 = (1 - \frac{T}{2}) \sqrt{\frac{n^2 - 1}{n^2 - 4}},$$

a T je von Neumannův poměr $T = \dfrac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$. Pokud jsou prvky výběru vzájemně nezávislé a platí nulová hypotéza H_0: $\rho_a = 0$, má veličina t_n Studentovo rozdělení s $(n+1)$ stupni volnosti. Alternativní hypotéza je H_A: $\rho_a \neq 0$. Platí-li, že pro případ $|t_n| > t_{1-\alpha/2}(n+1)$ je nutno nulovou hypotézu H_0 o nezávislosti prvků výběru na hladině významnosti α zamítnout.

2. předpoklad: **Výběr je homogenní**
Homogenní výběr znamená, že všechny jeho prvky x_i, $i = 1, ..., n$, pocházejí ze stejného rozdělení s konstantním rozptylem σ^2. K nehomogenitě naměřených dat dochází všude tam, kde se vyskytuje výrazná nestejnoměrnost měřených vlastností vzorků nebo se náhle mění podmínky experimentů. Speciálním případem jsou vybočující měření.

Nehomogenita může být způsobena také nevhodnou specifikací souboru. Pokud lze daný výběr rozdělit podle nějakých logických kritérií do několika podskupin, je možno zpracovat statisticky každou podskupinu zvlášť a pak na základě testů shody středních hodnot v podskupinách rozhodnout, zda je toto dělení významné. Omezíme se na případ, kdy se v datech vyskytují vybočující hodnoty. Tyto hodnoty se co do velikosti výrazně liší od ostatních a lze je běžně identifikovat v grafech průzkumové analýzy. Vybočující měření silně zkreslují odhady polohy a zejména rozptylu s^2, takže zcela znehodnocují další statistickou analýzu.

Problém vybočujících měření je velmi komplikovaný. Při jejich ověřování se používá řada idealizovaných předpokladů. Je nutné znát jejich předpokládaný počet, jejich rozdělení a rozdělení zbývajících prvků výběru. Navíc je třeba sestrojit model, podle kterého se vybočující měření chovají. Testování vybočujících měření bez doplňkových informací je proto málo spolehlivé.

Jednoduchá technika, kdy se pouze předpokládá, že "správná" data mají normální rozdělení, je *modifikace dolní vnitřní hradby* B_D^* a *horní vnitřní hradby* B_H^*,

$$B_D^* = \tilde{x}_{0.25} - K\,(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \;, \quad B_H^* = \tilde{x}_{0.75} + K\,(\tilde{x}_{0.75} - \tilde{x}_{0.25})$$

Parametr *K* se volí tak, aby pravděpodobnost *P(n, K)*, že z výběru velikosti *n* a pocházejícího z normálního rozdělení, nebude žádný prvek mimo vnitřní hradby $[B_D^*, B_H^*]$, byla dostatečně vysoká, např. 0.95.

Při volbě *P(n, K)* = 0.95 lze v rozmezí $8 \le n \le 100$ použít aproximace $K \approx 2.25 - 3.6/n$. Pro takto určený parametr *K* se všechny prvky výběru, ležící mimo hradby $[B_D^*, B_H^*]$ považují za vybočující. Výhodou je robustnost postupu. Není třeba znát počet vybočujících bodů ani jejich rozdělení a neprojevují se ani různé efekty "maskování".

3. předpoklad: **Rozdělení výběru je normální**

Na předpokladu normality je založena celá standardní statistická analýza dat. V *testu kombinace výběrové šikmosti a špičatosti* dle Jarque-Berra se užívá testovací kritérium

$$\chi^2_{\exp} = \frac{\hat{g}_1^2}{D(\hat{g}_1)} + \frac{[\hat{g}_2 - E(\hat{g}_2)]^2}{D(\hat{g}_2)}$$

kde jsou výběrová šikmost $\hat{g}_1$ a její rozptyl $D(\hat{g}_1)$, resp. výběrová špičatost $\hat{g}_2$, její střední hodnota $E(\hat{g}_2)$ a rozptyl $D(\hat{g}_2)$. Za předpokladu normality má veličina $\chi^2_{\exp}$ asymptoticky $\chi^2(2)$-rozdělení. Prokáže-li se proto, že $\chi^2_{\exp} > \chi^2_{1-\alpha}(2)$, je nutno hypotézu o normalitě rozdělení výběru zamítnout.

Střední hodnota výběru, pocházejícího z normálního rozdělení je $E(\hat{g}_1) = 0$. Pro asymptotický rozptyl tohoto odhadu platí

$$D(\hat{g}_1) \approx \frac{(n-2)}{(n+1)(n+3)}$$

Momentový odhad *špičatosti* g_2 je

$$\hat{g}_2 = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2}$$

Střední hodnota tohoto odhadu pro výběry pocházející z normálního rozdělení je

$$E(\hat{g}_2) = 3 - \frac{6}{n+1}$$

a pro asymptotický rozptyl tohoto odhadu platí

$$D(\hat{g}_2) \approx \frac{24\,n\,(n-2)\,(n-3)}{(n+1)^2\,(n+3)\,(n+5)} \;.$$

Při stanovení libovolného bodového odhadu parametru je třeba určit vždy i jeho rozptyl. K docílení stejné "přesnosti" odhadů je třeba při užití méně efektivního odhadu provést větší počet měření *n*. Například u dat pocházejících z normálního rozdělení se musí při použití mediánu $\tilde{x}_{0.5}$ provést 1.6krát více měření než při použití aritmetického průměru $\bar{x}$, aby se docílilo stejné přesnosti odhadu.

(a) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$:	10.012
Odhad rozptylu s^2:	5.223E-03
Odhad směrodatné odchylky s:	0.0723
Odhad šikmosti $\hat{g}_1$:	-0.04
Odhad špičatosti $\hat{g}_2$:	3.08
(b) Test normality: tabulkový kvantil $\chi^2_{1-\alpha}(2)$:	5.992
Odhad χ^2_{exp} statistiky:	0.112
Závěr: Předpoklad normality přijat na spočtené hladině významnosti $\alpha = 0.9456$.	
(c) Test nezávislosti: tabulkový kvantil $t_{1-\alpha/2}(n+1)$:	1.984
Odhad von Neumannovy statistiky t_n:	1.218
Závěr: Předpoklad nezávislosti přijat na spočtené hladině významnosti $\alpha = 0.113$.	
(d) Detekce odlehlých bodů: metodou modifikované vnitřní hradby	
Dolní vnitřní hradba B_D:	9.783
Horní vnitřní hradba B_H:	10.245
Závěr: Ve výběru nejsou odlehlé body.	

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. Často je pak vhodná transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě rozdělení.

1. *Stabilizace rozptylu* vyžaduje nalezení transformace $y = g(x)$, ve které je již rozptyl $\sigma^2(y)$ konstantní. Pokud je rozptyl původní proměnné x funkcí typu $\sigma^2(x) = f_1(x)$, lze rozptyl $\sigma^2(y)$ určit

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

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

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

2. *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 i *rankitového grafu*, kdy pro optimální exponent λ budou kvantily $y_{(i)}$ ležet na přímce nebo *selekčního grafu dle Hinese a Hinesové*.

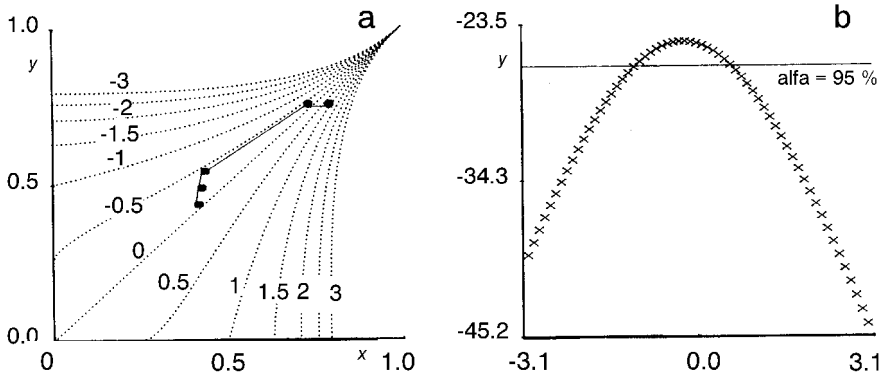
Hinesův - Hinesové selekční graf (osa x: $\tilde{x}_{0.5}/\tilde{x}_{1-P_i}$, osa y: $\tilde{x}_{P_i}/\tilde{x}_{0.5}$).
Diagnostickou pomůckou pro odhad optimálního exponentu λ je selekční graf dle Hinese a Hinesové. Vychází z požadavků symetrie jednotlivých kvantilů kolem mediánu

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

kde jako kvantily jsou obvykle voleny písmenové hodnoty. K porovnání průběhu jednotlivých bodů s ideálním pro zvolené λ se do grafu zakreslují řešení rovnice $y^{\lambda} + x^{-\lambda} = 2$ pro $0 \le x \le 1$ a $0 \le y \le 1$:

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

Podle umístění experimentálních bodů v okolí nomogramu teoretických křivek selekčního grafu lze vizuálně odhadovat velikost λ a posuzovat tak kvalitu transformace v různých vzdálenostech od mediánu.



Obr. 2.19 (a) Hinesův-Hinesové selekční graf a (b) graf logaritmu věrohodnostní funkce na λ pro výběr z lognormálního rozdělení, ***ADSTAT***

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 parametru 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 $L(\lambda)$ lze znázornit ve zvoleném intervalu např. $-3 \leq \lambda \leq 3$ a identifikovat maximum křivky, jehož x-ová souřadnice indikuje odhad $\hat{\lambda}$.

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 $\langle \lambda_D, \lambda_H \rangle$ širší, tím je mocninná Boxova-Coxova transformace méně výhodná. Pokud obsahuje interval $\langle \lambda_D, \lambda_H \rangle$ i hodnotu $\lambda = 1$, není transformace ze statistického hlediska přínosem.

Zpětná transformace: po vhodné transformaci vyčíslíme $\bar{y}$, $s^2(y)$ 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 vesměs vede k lepším odhadům polohy $\bar{x}_R$ a rozptylu $s^2(\bar{x}_R)$ a je vhodný zvláště v případech asymetrického rozdělení výběru.

Mocninná a Box-Coxova transformace u výběru *norm* a **úlohy B2.04** (ADSTAT)

(1) Odhady klasických parametrů:	*norm*	**B2.04**
Odhad aritmetického průměru	10.012	0.177
Odhad směrodatné odhylky	0.072271	0.159
Odhad šikmosti	-0.037	1.54
Odhad špičatosti	3.08	5.36
(2) Prostá mocninná transformace:		
Odhad optimálního exponentu	2.67	0.53

Odhad průměru transformovaných dat	465.69	0.355
Odhad směrodatné odchylky transformovaných dat	8.96	0.191
Odhad šikmosti transformovaných dat	0.0006	0.28
Odhad špičatosti transformovaných dat	3.08	3.13
Opravený odhad průměru původních dat	10.012	0.143
Opravený odhad směrodatné odchylky původních dat	0.07226	0.145
Spodní mez intervalu spolehlivosti původních dat	9.998	0.102
Horní mez mez intervalu spolehlivosti původních dat	10.027	0.190
(3) Box-Coxova transformace:		
Odhad optimálního exponentu	2.67	0.53
Odhad průměru transformovaných dat	174.26	-1.210
Odhad směrodatné odchylky transformovaných dat	3.361	0.358
Odhad šikmosti transformovaných dat	0.0006	0.28
Odhad špičatosti transformovaných dat	3.08	3.13
Opravený odhad průměru původních dat	10.012	0.143
Opravený odhad směrodatné odchylky původních dat	0.07226	0.145
Spodní mez intervalu spolehlivosti původních dat	9.998	0.102
Horní mez mez intervalu spolehlivosti původních dat	10.027	0.190

2.4 Průběh průzkumové analýzy dat

Průběh vlastní průzkumové, exploratorní analýzy dat (EDA) je možné libovolně kombinovat dle dosavadních informací o vyšetřovaných datech. Omezíme se na zpracování dvojího druhu dat, *rutinních dat*, o kterých jsou známy vlastnosti jako je např. rozdělení a jednak *neznámých dat*, o kterých nejsou známy dosud žádné předběžné informace a hrozí nebezpečí nesplnění předpokladů o datech.

A. Postup analýzy rutinních dat

Při zpracování rutinních výsledků měření předpokládáme, že známe rozdělení dat. 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 *rankitového grafu* a *grafu 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 řady grafických diagnostik,
b) určení výběrového rozdělení a jeho konstrukce.

Pokud nebylo nalezeno vhodné aproximující rozdělení, provádí se *mocninná transformace*, která by měla zlepšit rozdělení dat. Kombinace metod závisí na konkrétních datech a konkrétních požadavcích analýzy.

B. Postup při nesplnění předpokladů o datech

1. Nesplnění předpokladu nezávislosti prvků: Pokud prvky měření nejsou nezávislé, vzrůstá nebezpečí, že odhady budou systematicky vychýleny a nadhodnoceny pro pozitivní hodnotu autokorelačního koeficientu ρ_a. Nezbývá, než hlouběji analyzovat logické příčiny a snažit se o jejich odstranění, zkontrolovat celý měřicí řetězec a provést nová měření.

2. Nesplnění předpokladu normality výběru: Rozdělení dat je buď jiné než normální, nebo jsou v datech vybočující měření. V případě nenormálního rozdělení dat může jít o odchylky pouze v délce konců, nebo se jedná *o sešikmená rozdělení*. V případě symetrických rozdělení, lišících se od normálního délkou konců lze použít pro odhad parametrů polohy a rozptýlení jednoduché robustní techniky. U sešikmených rozdělení je vždy výhodné začít hledáním mocninné transformace. Pokud byla mocninná transformace úspěšná a byla nalezena optimální mocnina λ, provádí se další analýza v této transformaci a nakonec se vyčíslí zpětná transformace do původních proměnných.

Pro *sešikmená rozdělení*, charakterizovaná třetím centrálním momentem m_3,lze definovat modifikovanou náhodnou veličinu

$$T_C = \left[(\bar{x} - \mu) + \frac{m_3}{6\,\sigma^2\,n} + \frac{m_3}{3\,\sigma^4}\,(\bar{x} - \mu)^2 \right] \frac{\sqrt{n}}{s}$$

která má Studentovo rozdělení s (n - 1) stupni volnosti. Při praktických výpočtech se rozptyl σ^2 nahrazuje nevychýleným odhadem s^2 a třetí centrální moment m_3 jeho nevychýleným odhadem

$$\hat{m}_3 = \frac{n}{(n-1)\,(n-2)} \sum_{i=1}^{n} (x_i - \bar{x})^3$$

Při konstrukci *konfidenčního intervalu střední hodnoty* $L_D \le \mu \le L_H$ se užívá vztahů pro dolní a horní meze

$$L_D = \bar{x} + \frac{1}{2\,C_2} - \frac{\sqrt{d_1}}{2\,C_2}\,, \quad L_H = \bar{x} + \frac{1}{2\,C_2} - \frac{\sqrt{d_2}}{2\,C_2}$$

kde

$$C_1 = \frac{\hat{m}_3}{6\,s^2\,n} \qquad C_2 = \frac{\hat{m}_3}{3\,s^4}$$

$$C = t_{1-\alpha/2}(n-1)\,\frac{s}{\sqrt{n}}$$

$$d_1 = 1 - 4\,C_2\,(C_1 - C) \qquad d_2 = 1 - 4\,C_2\,(C_1 + C)$$

S využitím tohoto konfidenčního intervalu pro střední hodnotu sešikmených rozdělení lze také provádět testy významnosti parametru polohy.

3. Přítomnost vybočujících hodnot: Na základě logické analýzy je třeba nejdříve zvážit, zda nejde o sešikmené rozdělení. Body, které se jeví vybočující pro symetrické (speciálně normální) rozdělení, mohou být pro sešikmená rozdělení naopak přijatelné. Pokud jde o vybočující pozorování, lze použít dvou alternativ.

První alternativa spočívá v jejich vyloučení z další analýzy, což však není vždy zcela nejvhodnější. Pokud jsou vybočující měření výsledkem řídce se vyskytujících jevů, může tím totiž dojít ke ztrátě informace úplně. Proto lze tyto hodnoty vyloučit jedině při doplnění o nová experimentální data.

Druhá alternativa spočívá v použití robustních metod. Tento postup však nemusí být vždy korektní. Robustnost spočívá v přiblížení se k přijatému modelu měření bez ohledu na jeho platnost. Pokud se analýzy vybočujících měření účastní experimentátor, měl by rozhodnout, která měření jsou evidentní hrubé chyby (jako je selhání přístroje, špatný zápis dat), a která jsou jen podezřelá. Evidentní hrubé chyby je vhodné z další analýzy vyloučit, ale podezřelá měření je lépe ponechat. Robustními metodami se jejich vliv na odhady parametrů výrazně oslabí.

4. Nedostatečný rozsah výběru: Nejjednodušší je v tomto případě provést dodatečná měření. Platí, že čím jsou data méně rozptýlená, tím menší počet jich stačí k zajištění dostatečné přesnosti odhadu. Pokud nelze provést dodatečné experimenty, je možné použít techniky vhodné pro malé výběry (viz Hornův postup ve 3. kap.).

Tento postup je vhodný zejména pro analýzu rutinních měření, kde jsou o chování dat předběžné informace. Když se analyzují výsledky nových měření nebo neznámé výběry, je vždy třeba začít průzkumovou analýzou dat a stanovit statistické zvláštnosti výběru.