Když se řekne...

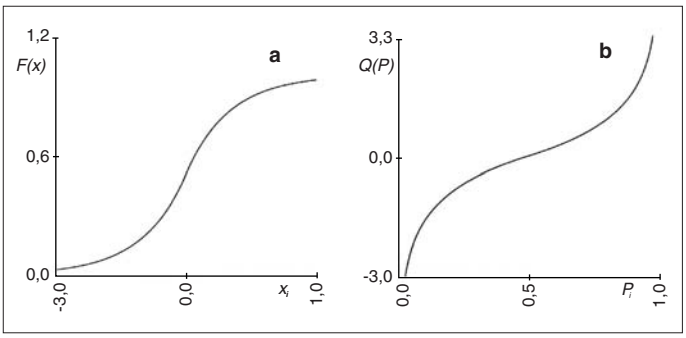
Interaktivní statistická analýza dat

V posouzení správnosti naměřených výsledků tvoří průzkumová (exploratorní) analýza dat důležitou pomůcku. Využívá kvantilových grafických diagnostik ke sledování stupně symetrie a špičatosti rozdělení výběru, lokální koncentrace dat a přítomnosti vybočujících hodnot. Mezi nejdůležitější patří vedle kvantilového grafu a grafu rozptýlení s kvantily také krabicový graf, vrubový krabicový graf, graf polosum a graf symetrie, kvantilový graf, rankitový graf, jádrový odhad hustoty pravděpodobnosti a histogram. Intervalový odhad míry polohy a Studentův t-test správnosti jsou hlavními testy k posouzení správnosti. U malých výběrů $4 \le n \le 20$ je výhodný Hornův postup pivotů, který je vedle malé četnosti vhodný pro svou dostatečnou robustnost vůči asymetrii rozdělení a vůči vybočujícím hodnotám.

1 Úvod

Otázka spolehlivosti a správného vyhodnocení experimentálních dat se v době osobních počítačů ocitá u každého měření dat na prvním místě. V kontrolní laboratoři, ať už chemické, biologické, fyzikální či jakékoliv jiné, tvoří základ experimentální práce měření na přístroji. V laboratořích dnes představují instrumentální metody spojovací článek mezi přírodovědnými a technickými obory, protože moderní přístroje s vestavěným procesorem používá každá laboratoř. Na každém psacím stole v laboratoři nacházíme osobní počítač, většinou nejvyšší kvality, kapacity a rychlosti a vybavený moderním softwarem. Je proto poněkud neomluvitelné vyhodnocovat naměřená data zjednodušenými, aproximativními postupy z doby kalkulaček. Kontrolní orgány, komisaři akreditačních komisí, ale především konkurenční pracoviště v zahraničí užívají k vyhodnocení dat špičkový software s rigorózními matematickými postupy, ve kterých není žádné zjednodušení či zanedbání důležitých statistických předpokladů a výsledky získané těmito náročnějšími postupy jsou považovány za platné (validní) a správné, přijatelné třeba v okružním testu.

Ukažme si jeden z novějších postupů *interaktivní statistické analýzy dat,* který je založen na diagnostikování uživatele v dialogu s osobním počítačem čili na interaktivní analýze, který nabízí hlubší pohled do všech tajemství ukrytých v datech. S problémem souvisí obvykle i vhodný software, který zajistí bezproblémové a přátelské prostředí a „nechá naše data promluvit". Nezapomeňme přitom na důležité pravidlo, že *úroveň užívaného softwaru dnes prozrazuje úroveň pracoviště.*



Obr. 1 (a) Distribuční funkce $F(x_i)$ a (b) kvantilová funkce $Q(P_i)$ Laplaceova rozdělení s nulovou střední hodnotou a rozptylem rovným 2

2 Postup interaktivní analýzy dat

Interaktivní přístup ulehčuje postup interaktivní analýzy dat, protože většina statistického softwaru obsahuje uvedené statistické diagnostiky a testy. Obecný postup náročnější statistické analýzy jednorozměrných dat spočívá v níže uvedených krocích:

1. *Průzkumová (exploratorní) analýza dat (EDA)* vyšetřuje data s cílem určit stupeň symetrie a špičatosti rozdělení, lokální koncentrace dat a rozdělení výběru a odhalení vybočujících a podezřelých dat.
2. *Ověření předpokladů výběru dat* se týká ověření normality, ověření nezávislosti, ověření homogenity a konečně i určení minimální četnosti.
3. *Transformace dat* následuje v případě porušení některého z předpokladu o výběru. Patří sem mocninná transformace a Boxova-Coxova transformace.
4. *Vyčíslení nejlepších odhadů parametrů polohy, rozptýlení a tvaru* se týká vyčíslení jednak klasických odhadů (aritmetický průměr a rozptyl), jednak robustních odhadů (medián, uřezané průměry, „winsorizovaný" rozptyl) a konečně i adaptivních *M* odhadů.
5. *Testování výběrů* se týká obvykle testů správnosti a testů shodnosti.

3 Diagnostiky v interaktivní analýze dat

3.1 Základní pojmy

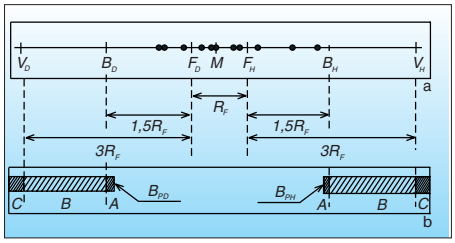
Prvním krokem v analýze jednorozměrných dat je průzkumová, exploratorní analýza. Jejím cílem je odhalit statistické zvláštnosti v datech a ověřit předpoklady o výběru pro následné rigorózní statistické zpracování. Jedině tak lze zabránit provádění numerických výpočtů bez hlubších statistických souvislostí. Z různých typů výběru se v laboratoři nejvíce uplatňuje reprezentativní náhodný výběr, $\{x_i\}$, $i = 1, ..., n$, který má následující vlastnosti:

- jednotlivé prvky výběru x_i jsou vzájemně nezávislé,
- výběr je homogenní, tj. všechna x_i pocházejí ze stejného rozdělení pravděpodobnosti s konstantním rozptylem,
- jde o normální rozdělení pravděpodobnosti,
- všechny prvky souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru.

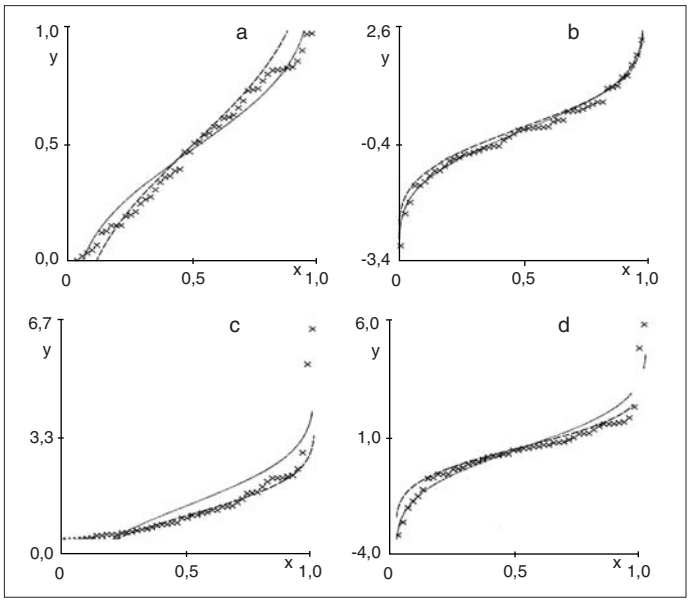
Před vlastní analýzou je vždy nezbytné ověřit platnost základních předpokladů, tj. nezávislost, homogenitu a normalitu výběru. Využívá se především robustních kvantilových charakteristik, které umožňují sledování lokálního chování dat a které jsou vhodné pro malé nebo střední velké výběry. Vychází se z *pořádkových statistik* výběru $x_{(1)} \le x_{(2)} ... \le x_{(n)}$. Platí, že střední hodnota *i*-té pořádkové statistiky je rovna $100P_i$ procentnímu kvantilu výběrového rozdělení $F^{-1}(P_i) = Q(P_i)$, kde $F(x)$ označuje distribuční funkci a $Q(P_i)$ kvantilovou funkci výběru. Symbol $P_i = i/(n + 1)$ označuje *pořadovou pravděpodobnost*. Připomeňme, že $100P_i$ *procentní výběrový kvantil* je hodnota, pod kterou leží $100P_i$ procent prvků výběru. Optimální hodnoty P_i závisí na předpokládaném rozdělení výběru. Pro normální rozdělení se často doporučuje volba $P_i = (i - 3/8)/(n + 1/4)$. Vynesením hodnot $x_{(i)}$ proti P_i, $i = 1, ..., n$, se získá hrubý odhad *kvantilové funkce* $Q(P)$. Ta je inverzní k funkci distribuční a jednoznačně charakterizuje rozdělení výběru (*obr. 1*). V průzku-

Tab. 1 Označení písmenových hodnot

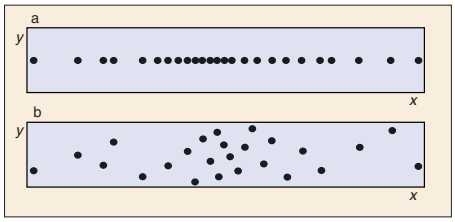
i	*i*-tý kvantil	Pořadová pravděpodobnost P_i	Symbol písmenové hodnoty *L*	Hodnota kvantilu u_{Pi}
1	medián	$2^{-1} = 1/2$	*M*	0
2	kvartily	$2^{-2} = 1/4$	*F*	–0,674
3	oktily	$2^{-3} = 1/8$	*E*	–1,15
4	sedecily	$2^{-4} = 1/16$	*D*	–1,53



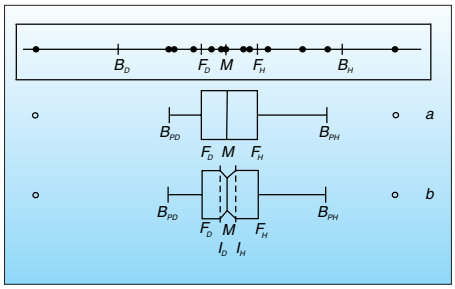
Obr. 2 Konstrukce bariérově-číslicového schématu indikujícího vybočující hodnoty: a) diagram rozptýlení s mediánem *M*, kvartily F_D (dolní) a F_H (horní), vnitřní hradby B_D (dolní) a B_H (horní), vnější hradby V_D (dolní) a V_H (horní); b) oblast vybočujících hodnot: A přilehlé (B_{PD} je blízké B_D a B_{PH} je blízké B_H), B značí oblast vnějších a C vzdálených bodů



Obr. 3 Kvantilové grafy (robustní -- a klasické ..) pro výběry z rozdělení a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova



Obr. 4 Konstrukce a) diagramu rozptýlení a b) rozmítnutého diagramu rozptýlení



Obr. 5 Konstrukce a) krabicového grafu, a b) vrubového krabicového grafu z dat diagramu rozptýlení; prázdné kroužky indikují vybočující hodnoty

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

Symbol u_{Pi} 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 požadovanou 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 $R_{Pi} = i$ a klesající pořadí $K_{Pi} = n + 1 - i$. Hloubka H_i je pak menší číslo z obou pořadí $H_i = \min(R_{Pi}\, K_{Pi})$.

Na *obr. 2* je znázorněna konstrukce bariérově-číslicového schématu indikujícího vybočující hodnoty. Horní část obrázku (a) znázorňuje diagram rozptýlení s mediánem *M*, kvartily F_D (dolní) a F_H (horní), vnitřními hradbami B_D (dolní) a B_H (horní) a vnějšími hradbami V_D (dolní) a V_H (horní). V dolní části (b) je zakreslena oblast vybočujících hodnot: písmenem A jsou označeny oblasti přilehlých bodů, písmeno B (B_{PD} je blízké B_D a B_{PH} je blízké B_H), značí oblast vnějších a C vzdálených bodů.

Pro hloubku mediánu platí $H_M = (n + 1)/2$. Pokud je tato hloubka celé číslo, je medián $x_{0,5} = M = x_{(HM)}$. V opačném případě se provádí lineární interpolace mezi $x_{(n/2)}$ a $x_{(n/2+1)}$. Hloubky dolních písmenných hodnot jsou $H_L = (1 + \text{int}(H_{L-1}))/2$, kde L jsou indexy F, E, D a int(x) značí celočíselnou část čísla x. Pokud je $L = F$, bere se $L - 1 = M$. Jestliže je H_L celé číslo, bude dolní kvantilů $L_D = x_{(HL)}$ a horní kvantilů $L_H = x_{(n+1-HL)}$. Jeli H_L číslo necelé, provádí se lineární interpolace. Tento postup se pro menší hodnoty H_L, kdy jsou kvantilů 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 $n_L = 1{,}44 \ln(n + 1)$.

3.2 Kvantilový graf

V kvantilovém grafu (*obr. 3*) se na ose x vynáší pořadová pravděpodobnost P_i a na ose y pořádková statistika $x_{(i)}$. Graf umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, který může být symetrický, zešikmený k vyšším nebo nižším hodnotám. Ke snadnějšímu porovnání s normálním rozdělením se do tohoto grafu zakreslují i kvantilové funkce normálního rozdělení $N_{Pi} = \hat{\mu} + \hat{\sigma}\, u_{Pi}$, pro $0 \le P_i \le 1$, a to jak u klasických odhadů parametrů polohy a rozptýlení, kde $\hat{\mu} = \bar{x}$ a $\sigma = s$, tak i u odhadů robustních, kde $\hat{\mu} = \tilde{x}_{0,5}$ a $\sigma = R_F/1{,}349$.

3.3 Diagram rozptýlení

Diagram rozptýlení (*obr. 4a*) představuje jednorozměrnou projekci kvantilového grafu do osy x. I při své jednoduchosti tento diagram ukazuje na lokální koncentrace dat a indikuje i podezřelá a vybočující měření. Projekci kvantilového grafu představuje také *rozmítnutý diagram rozptýlení* (*obr. 4b*), v němž se na osu y vynese interval náhodných čísel, a tím se data příhodně rozmítnou.

3.4 Krabicový graf

V krabicovém grafu se na osu x vynesou hodnoty úměrné hodnotám x a osa y představuje libovolný interval. Graf představuje obdélník o délce $R_F = F_H - F_D = \tilde{x}_{0,75} - \tilde{x}_{0,25}$ s vhodně zvolenou šířkou, která je úměrná hodnotě n. V místě mediánu je vertikální čára. Od obou protilehlých stran tohoto obdélníku pokračují úsečky. Ty jsou ukončeny *přilehlými hodnotami* B_{PH} a B_{PD}, ležícími uvnitř *vnitřních hradeb* nejblíže k jejich hranicím B_H, B_D, tj. $B_H = F_H + 1{,}5\, R_F$ a $B_D = F_D - 1{,}5\, R_F$. Pro data pocházející z normálního rozdělení platí $B_H - B_D = 4{,}2$. Prvky výběru mimo vnitřní hradby jsou považovány za podezřelá měření (kroužky).

Graf slouží pro částečnou sumarizaci dat, a to pomocí následujících charakteristik (*obr. 5a*):

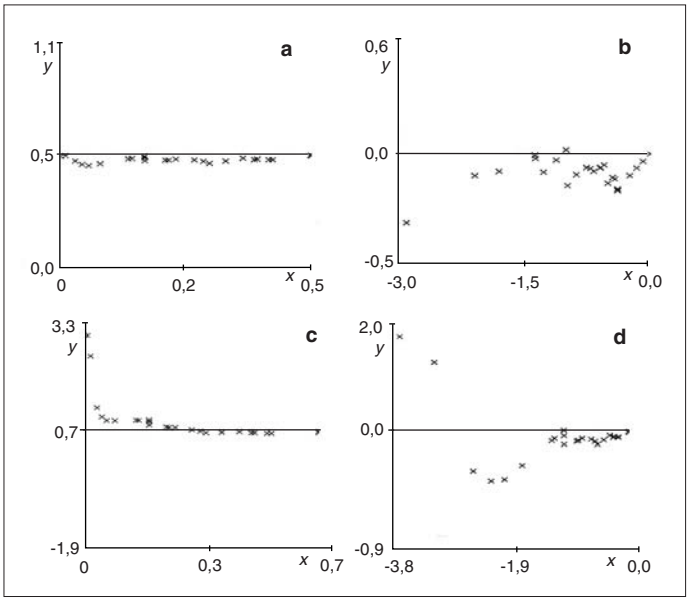
- znázornění robustního odhadu polohy mediánu *M*,
- posouzení symetrie v okolí kvantilů,
- posouzení symetrie u konců rozdělení,
- identifikaci odlehlých dat.

Obdobou krabicového grafu je *vrubový krabicový graf* (osa x: úměrná hodnotám x, osa y: libovolný interval), který umožňuje i posouzení variability mediánu. Ta je vyjádřena robustním intervalem spolehlivosti $I_D \le M \le I_H$ (*obr. 5b*).

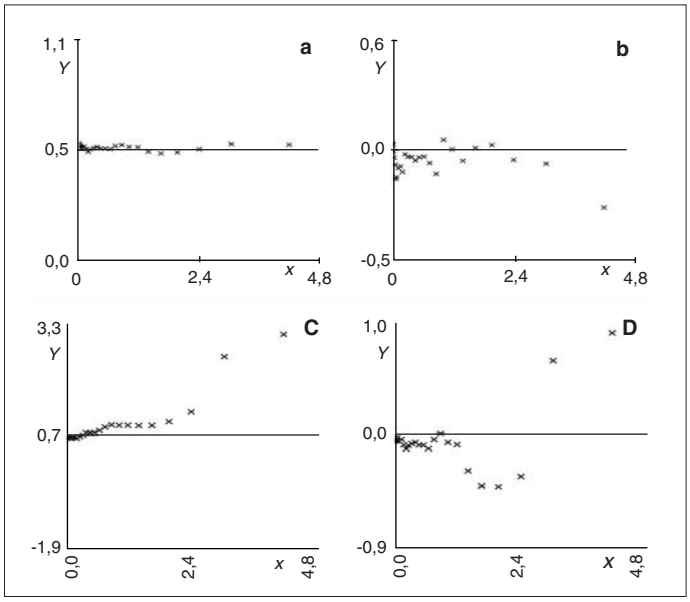
3.5 Graf polosum a graf symetrie

V grafu polosum (*obr. 6*) se na osu x vynáší pořádkové statistiky $x_{(i)}$ na osu y hodnoty polosum dané vztahem: $Z_i = 0{,}5\,(x_{(n+1-i)} + x_{(i)})$. Pro symetrické rozdělení je grafem horizontální přímka, určená rovnicí $x_{0,5} = M$.

Graf symetrie (*obr. 7*) má na ose x hodnoty odpovídající vztahu $u_{Pi}^2/2$ pro $P_i = i/(n + 1)$ a na ose y hodnoty podle vztahu $Z_i = 0{,}5(x_{(n+1-i)} + x_{(i)})$. V tomto grafu jsou symetrická rozdělení charakterizována hori-



Obr. 6 Grafy polosum pro výběry z rozdělení a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova



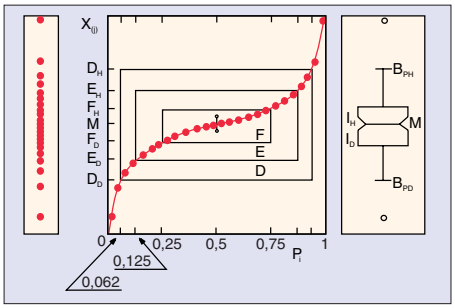
Obr. 7 Grafy symetrie pro výběry z rozdělení a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova (— značí symetrii)

zontální přímkou $\tilde{y} = x_{0,5} = M$. Pokud tato přímka nemá nulovou směrnici, je směrnice odhadem parametru šikmosti.

3.6 Graf rozptýlení s kvantily

Základem grafu rozptýlení s kvantily (osa *x*: P_i, osa *y*: $x_{(i)}$) je odhad kvantilové funkce výběru, který se získá spojením bodů $\{x_{(i)}, P_i\}$ lineárními úseky. Konstrukce grafu je znázorněna na *obr. 8*. Pro srovnání je svisle vlevo umístěn diagram rozptýlení a vpravo vrubový krabicový graf, ve kterém prázdná kolečka indikují vybočující hodnoty Pro symetrická rozdělení má kvantilová funkce sigmoidální tvar. Pro rozdělení zešikmená k vyšším hodnotám je konvexně rostoucí a pro rozdělení zešikmená k nižším hodnotám konkávně rostoucí. Do grafu se zakreslují tři obdélníky *F, E* a *D*:

- *kvartilový obdélník F:* na ose *x* jsou vyneseny pravděpodobnosti $P_2 = 2^{-2} = 0{,}25$ a $1 - 2^{-2} = 0{,}75$,
- *oktilový obdélník E*: na ose *y* je tvořen oktily E_D a E_H a na ose *x* pravděpodobnostmi $P_3 = 2^{-3} = 0{,}125$ a $1 - 2^{-3} = 0{,}875$,
- *sedecilový obdélník D*: na *y* jsou vyneseny sedecily D_D, D_H a na *x* pravděpodobnosti $P_4 = 2^{-4} = 0{,}0625$ a $1 - 2^{-4} = 0{,}9375$.



Obr. 8 Konstrukce grafu rozptýlení s kvantily

Podle této grafické diagnostiky lze určit následující charakteristiky výběru:

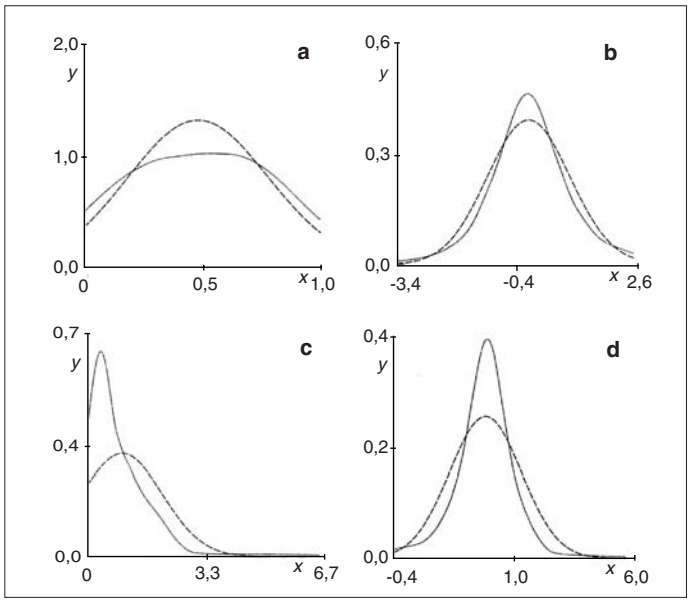
- *symetrické unimodální rozdělení* výběru obsahuje obdélníky symetricky uvnitř sebe,
- *nesymetrická rozdělení* mají pro rozdělení zešikmené k vyšším hodnotám vzdálenosti mezi dolními hranami obdélníků *F, E* a *D* výrazně kratší než mezi jejich horními hranami,
- *odlehlá pozorování* jsou indikována tím, že na kvantilové funkci mimo obdélník F se objeví náhlý vzrůst, kdy hodnota směrnice roste nade všechny meze,
- *vícemodální rozdělení* jsou indikována tím, že na kvantilové funkci uvnitř obdélníku *F* je několik úseků s téměř nulovými směrnicemi.

3.7 Jádrový odhad hustoty pravděpodobnosti

Jádrový odhad hustoty pravděpodobnosti se konstruuje tak, že se na osu *x* vynesou hodnoty *x* a na osu *y* příslušné hustoty pravděpodobnosti. Na *obr. 9* jsou tyto grafy pro výběry dat z různých typů rozdělení. Čárkováním je znázorněna hustota Gaussova rozdělení s parametry $\bar{x}$ a s^2 a plnou čarou jádrový odhad hustoty pravděpodobnosti empirického rozdělení výběru.

3.8 Histogram

Jedním z nejstarších klasických odhadů hustoty pravděpodobnosti je *histogram* (osa *x*: proměnná *x*, osa *y*: hustota pravděpodobnosti) – *obr. 10*. 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 roz-



Obr. 9 Jádrové odhady hustoty pravděpodobnosti pro výběry z rozdělení (a) rovnoměrného, (b) normálního, (c) exponenciálního a (d) Laplaceova. čárkováním je znázorněna hustota Gaussova rozdělení s parametry $\bar{x}$ a s^2 a plnou čarou jádrový odhad hustoty pravděpodobnosti empirického rozdělení výběru

dělení výběru lze počítat L podle vztahu

$$L = \mathrm{int}(2\sqrt{n}) \qquad (1)$$

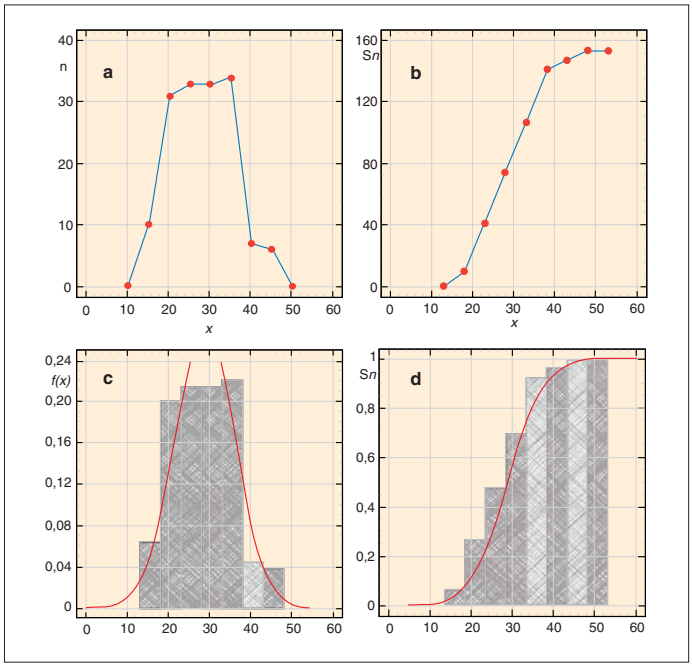
kde funkce int(x) označuje celočíselnou část čísla x. V širokém rozmezí velikostí výběrů n je možné užít výraz $L = \mathrm{int}\,(2{,}46\,(n-1)^{0{,}4})$.

3.9 Kvantil-kvantilový graf (graf Q-Q)

Graf Q-Q (osa x: $Q_T(P_i)$, osa y: $x_{(i)}$) uvedený na *obr. 11* umožňuje 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 užívají pořádkové statistiky $x_{(i)}$. Při shodě výběrového rozdělení se zvoleným teoretickým rozdělením musí platit přibližná rovnost kvantilů $x_{(i)} = Q_T(P_i)$, kde P_i je pořadová pravděpodobnost. Pokud je rozdělení výběru shodné se zvoleným teoretickým rozdělením, je závislost $x_{(i)}$ na $Q_T(P_i)$ lineární. Tato závislost se nazývá graf *Q-Q*.

3.10 Rankitový graf

Graf Q-Q pro porovnání rozdělení výběru s rozdělením normálním se nazývá *rankitový graf* (osa x: kvantil normovaného normálního rozdělení u_{Pi}, osa y: $x_{(i)}$) uvedený na *obr. 12*. Umožňuje orientačně zařadit rozdělení výběru do skupin podle šikmosti, špičatosti a délky konců.



Obr. 10 Histogram a) s grafem hustoty pravděpodobnosti – čárkovaně, b) kumulativní histogram četností, c) polygon četností a d) polygon kumulativních četností

4 Intervalový odhad parametrů

4.1 Interval spolehlivosti

Odhadem se stanoví interval, ve kterém se bude se zadanou pravděpodobností se statistickou jistotou $(1-\alpha)$ nacházet skutečná hodnota čili „pravda“ daného parametru Θ. Neznámý parametr Θ odhadujeme dvěma číselnými hodnotami L_D a L_H, které tvoří meze tzv. *intervalu spolehlivosti* čili konfidenčního intervalu. Interval spolehlivosti pokryje parametr Θ s předem zvolenou, statistickou jistotou čili dostatečně velikou pravděpodobností $P = (1-\alpha)$, což lze vyjádřit vztahem $P(L_D < \Theta < L_H) = 1-\alpha$, nazvanou *koeficient spolehlivosti* (čili konfidenční koeficient, statistická jistota). Je obyčejně roven 0,95 nebo 0,99. Parametr α se nazývá *hladina významnosti*. Interval spolehlivosti vyjadřuje tvrzení: „*Statistická jistota, s jakou bude ‚pravda‘ Θ ležet v náhodných mezích L_D, L_H je rovna právě $1-\alpha$.*“ Interval spolehlivosti má následující vlastnosti:

- čím je rozsah výběru n větší, tím je interval spolehlivosti užší,
- čím je odhad přesnější a má menší rozptyl, tím je interval spolehlivosti užší,
- čím je vyšší statistická jistota $(1-\alpha)$, tím je interval spolehlivosti širší.

4.2 Konstrukce intervalových odhadů

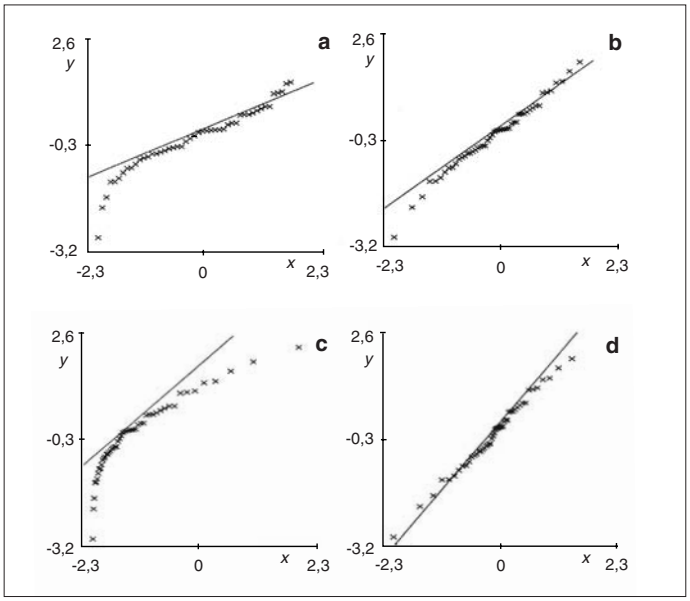
Postup konstrukce intervalu spolehlivosti střední hodnoty μ normálního rozdělení $N(\mu, \sigma^2)$ se liší podle velikosti výběru:

Pro *velký výběr* $n \geq 30$ platí, že je-li nejlepším bodovým odhadem střední hodnoty μ výběrový průměr $\bar{x}$ s rozdělením $N(\mu, \sigma^2/n)$, pak v intervalu $\bar{x} \pm 1{,}96\sigma/\sqrt{n}$ leží přibližně 95 % hodnot náhodných veličin výběru o rozsahu n a *$100(1-\alpha)$% interval spolehlivosti střední hodnoty* μ bude vyčíslen podle vztahu

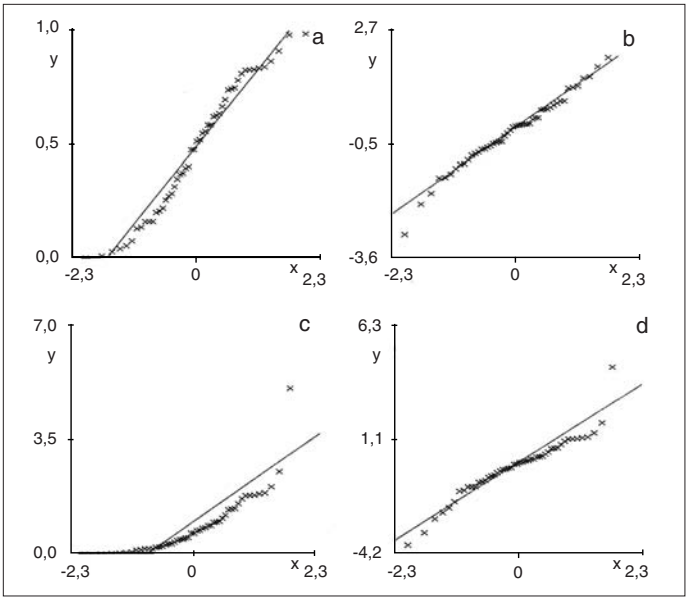
$$\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}} \qquad (2)$$

kde hodnota 1,96 je $100(1-0{,}05/2) = 97{,}5\%$ kvantil normovaného normálního rozdělení $u_{0{,}975}$.

Pro *malý výběr* $n \leq 30$ v praxi obvykle



Obr. 11 Grafy *Q-Q* pro porovnání rozdělení výběru normálního rozdělení s rozdělením a) teoretickým rovnoměrným, b) normálním, c) exponenciálním a d) Laplaceovým; čarou je vyznačen teoretický průběh



Obr. 12 Rankitové grafy pro výběry z rozdělení a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova; plnou čarou je vyznačen teoretický průběh

neznáme směrodatnou odchylku , ale pouze její odhad s a pokud je $100(1-\alpha/2)\%$ kvantil Studentova rozdělení roven $t_{1-\alpha/2}(n-1)$, bude $100(1-\alpha)\%$ *interval spolehlivosti střední hodnoty* μ roven

$$\bar{x} - t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}} \quad (3)$$

Meze intervalu spolehlivosti závisí vedle chyby s i na rozsahu výběru n. Pro větší rozsahy výběru ($n > 30$) lze použít místo kvantilu $t_{1-\alpha/2}$ kvantilu normovaného normálního rozdělení $u_{1-\alpha/2}$.

Obecně proto platí: $100(1-\alpha)\%$ *interval spolehlivosti parametru polohy* Θ se vypočte podle asymptotického vztahu

$$\hat{\Theta} - u_{1-\alpha/2}\sqrt{D(\hat{\Theta})} \le \Theta \le \hat{\Theta} + u_{1-\alpha/2}\sqrt{D(\hat{\Theta})} \quad (4)$$

a $100(1-\alpha)\%$ *oboustranný interval spolehlivosti rozptylu* σ^2 se vypočte podle vztahu

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \quad (5)$$

kde $\chi^2_{1-\alpha/2}(n-1$je horní a $\chi^2_{\alpha/2}(n-1)$ dolní kvantil rozdělení χ^2.

Robustní *interval spolehlivosti mediánu* se přibližně vyčíslí takto:

$$\tilde{x}_{0,5} - u_{1-\alpha/2}\frac{0,707s}{\sqrt{n}} \le med$$

$$med \le \tilde{x}_{0,5} + u_{1-\alpha/2}\frac{0,707s}{\sqrt{n}} \quad (6)$$

5 Analýza malých výběrů

Předem je třeba si uvědomit, že závěry z malých výběrů jsou vždy zatíženy značnou mírou nejistoty. Malé rozsahy lze proto využít jen tam, kde skutečně není možné zvýšit počet měření.

Pro *zvláště malé výběry* o $n = 2$ se $100(1-\alpha)\%$ *konfidenční interval střední hodnoty* se vyčíslí podle vztahu

$$\frac{x_1+x_2}{2} - T_\alpha\frac{|x_1-x_2|}{2} \le \mu \le \frac{x_1+x_2}{2} + T_\alpha\frac{|x_1-x_2|}{2} \quad (7)$$

Pro normální rozdělení bude $T_\alpha = \mathrm{cotg}(\alpha\pi/2)$, $T_{0,05} = 12,71$ a pro rovnoměrné rozdělení $T_\alpha = 1/\alpha - 1$, tj. $T_{0,05} = 19$.

Pro zvláště malé výběry o $n = 3$ se $100(1-\alpha)\%$ *konfidenční interval střední hodnoty* vyčíslí podle vztahu

$$\bar{x} - T'_\alpha\frac{s}{\sqrt{3}} \le \mu \le \bar{x} + T'_\alpha\frac{s}{\sqrt{3}} \quad (8)$$

Pro normální rozdělení bude platit:

$$T'_\alpha \approx 1/\sqrt{\alpha} - 3\sqrt{\alpha}/4$$

Z tohoto vztahu se vyčíslí $T_\alpha = 4,30$. Pro rovnoměrné rozdělení je $T_{0,05} = 5,74$.

Hornův postup pro malé výběry $4 \le n \le 20$ je založený na pořádkových statistikách. Nejprve se určí hloubka pivotu podle vztahu $H = (\mathrm{int}((n+1)/2))/2$ nebo $H = (\mathrm{int}((n+1)/2 + 1)/2$, pak dolní pivot jako $x_D = x_{(H)}$ a horní pivot dle $x_H = x_{(n+1-H)}$. Odhadem parametru polohy je potom *pivotová polosuma* $P_L = (x_D + x_H)/2$ a odhadem parametru rozptýlení je *pivotové rozpětí* $R_L = x_H - x_D$. Lze definovat i náhodnou veličinu k testování $T_L = P_L/R_L$, která má přibližně symetrické rozdělení, jehož vybrané kvantily jsou v tabulkách, např. v [1]. *95% interval spolehlivosti střední hodnoty* se vypočte vztahem $P_L - R_L\, t_{L,0,975}(n) \le \mu \le P_L - R_L\, t_{L,0,975}(n)$.

6 Ilustrativní příklad

6.1 Test správnosti koncentrace tenzidů

Standardní vzorek obsahuje 2,5 mg/l aniontaktivních tenzidů. Je třeba testovat, zda výsledky koncentrace standardu jsou správné. Byly zjištěny následující koncentrace tenzidů [mg/l]: 2,36, 2,40, 2,4, 2,50, 2,57, 2,62, 2,68.

Nejprve se vyčíslí pořádkové statistiky koncentrace tenzidů a uspořádají se vzestupně podle uvedených hodnot (*tab. 2*). Pro výpočet hloubky pivotu použijeme pro tento výběr s lichým počtem prvků $n = 7$ vzorec:

$$H = \mathrm{int}\frac{\frac{n+1}{2}}{2}, \qquad \mathrm{int}(2,0) \approx 2$$

Pro dolní pivot platí $x_D = x_{(H)}$ a po dosazení bude $x_{(2)} = 2,40$. Pro horní pivot platí $x_H = x_{(n+1-H)}$ a po dosazení bude $x_{(6)} = 2,62$. Pilotová polosuma vychází $P_L = (x_D + x_H)/2 = 2,51$ a pivotové rozpětí $R_L = x_H - x_D = 2,62 - 2,40 = 0,22$. 95% interval spolehlivosti střední hodnoty μ: $t_{L,\,1-\alpha/2}(7) = 0,720$ a dosazením do vztahu

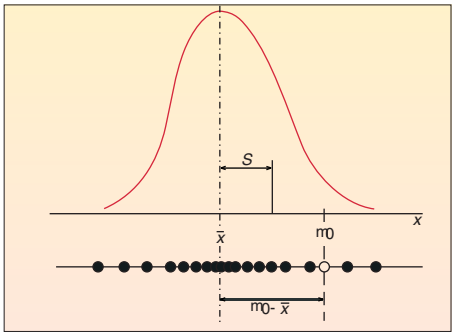
$$P_{L-}R_L t_{L,1-\alpha/2}(n) \le \mu \le P_{L\,i+}R_L t_{L,\,1-\alpha/2}(n)$$

vyjde $2,51 - 0,22 \times 0,72 \le \mu \le 2,51 + 0,22 \times 0,72$ a výsledkem bude nerovnost $2,35 \le \mu \le 2,67$

7 Testování statistických hypotéz

Testování statistické hypotézy se provádí podle následujícího obecného postupu:

1. Formulace nulové H_0 hypotézy a alternativní hypotézy H_A.
2. Volba hladiny významnosti α.
3. Volba testační statistiky, např. t.
4. Určení kritického oboru testové charakteristiky, např. $t_{1-\alpha/2}(n-1)$,



Obr. 13 Test správnosti výsledku $\bar{x}$ vůči normě μ_0

5. Vyčíslení testační statistiky a jejích kvantilů.
6. Rozhodnutí, zda (a) zamítnout hypotézu H_0 a přijmout H_A, jestliže testační statistika padne do kritického oboru, (b) nezamítnout hypotézu H_0, jestliže testační statistika nepadne do kritického oboru.

8 Test správnosti výsledku

Testy hypotéz (*obr. 13*) o parametrech μ a σ^2 při normálním rozdělení se provádí tak, že se ze souboru s $N(\mu, \sigma^2)$ s výběrem rozsahu n vypočte průměr x a směrodatná odchylka s. Hypotézu H_0 formulujeme takto: $\mu = \mu_0$ a hypotézu H_A takto: $\mu \ne \mu_0$ Testová statistika je dána vztahem:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n} \quad (9)$$

Dalším krokem je testování střední hodnoty μ a rozptylu σ^2. Pro výběr normálního rozdělení je $t_\alpha(n-1)$ kvantil Studentova a $\chi^2{}_\alpha(n-1)$ je kvantil χ^2-rozdělení (*tab. 3*). Hypotézu H_0 formulujeme takto $\sigma^2 = \sigma_0^2$ a hypotézu H_A takto: $\sigma^2 \ne \sigma_0^2$. Testová statistika je dána vtahem:

$$\chi^2 = \frac{(n-1)\,s^2}{\sigma_0^2} \quad (10)$$

čím je hodnota $(1-\alpha/2)$ u oboustranného testu bližší jedné (např. větší než 0,975), tím vě-

Tab. 2 Pořádkové statistiky koncentrace tenzidů uspořádané vzestupně

i	1	2	3	4	5	6	7
$x_{(i)}$	2,36	2,40	2,48	2,50	2,57	2,62	2,68

Tab. 3 Testování střední hodnoty μ a rozptylu σ^2 normálního rozdělení, kde $t_\alpha(n-1)$ je kvantil Studentova rozdělení

Nulová hypotéza	Alternativní hypotéza	Testační charakteristika	Kritický obor
$\mu = \mu_0$	$\mu > \mu_0$	$t = (x-\mu_0)\sqrt{n}/s$	$t \ge t_{(1-\alpha)}(n-1)$
	$\mu < \mu_0$		$t < t_\alpha(n-1)$
	$\mu \ne \mu_0$		$t \ge t_{(1-\alpha/2)}(n-1)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 = (n-1)s^2/\sigma_0^2$	$\chi^2 \ge \chi^2_{1-\alpha}(n-1)$
	$\sigma^2 < \sigma_0^2$		$\chi^2 < \chi^2{}_\alpha(n-1)$
	$\sigma^2 \ne \sigma_0^2$		$\chi^2_{\alpha/2}(n-1) \le \chi^2 \le \chi^2_{1-\alpha/2}(n-1)$

rohodnější bude zamítnutí nulové hypotézy H_0. Testy správnosti výsledku měření lze provést také pomocí intervalu spolehlivosti podle pravidla: pokud $100(1-\alpha)\%$ interval spolehlivosti parametru μ obsahuje zadanou hodnotu μ_0, nelze na hladině významnosti α zamítnout hypotézu $H_0 : \mu = \mu_0$.

9 Závěr

V postupu statistického vyhodnocení výsledků měření slouží průzkumová analýza dat EDA jako výhodná pomůcka k vyšetření zvláštností statistického chování dat. Z nejdůležitějších pomůcek se vedle kvantilového grafu a grafu rozptýlení s kvantilů používá i diagram rozptýlení a rozmítnutý diagram rozptýlení, krabicový graf, vrubový krabicový graf, graf polosum a symetrie, kvantilový graf, rankitový graf, jádrový odhad hustoty pravděpodobnosti a histogram k určení tvaru rozdělení.

U malých výběrů $4 \le n \le 20$ poskytuje správné odhady střední hodnoty Hornův postup pivotů. Pivotová polosuma a pivotové rozpětí umožňují vyčíslit intervalový odhad střední hodnoty a navíc jsou oba odhady dostatečně robustní vůči asymetrii rozdělení malého výběru a i vůči odlehlým hodnotám.

Studentův *t*-test správnosti analytického výsledku je rovnocenným testem vůči intervalu spolehlivosti. Nachází-li se totiž hodnota μ_0 (tj. „pravda“, správná hodnota, norma, standard) v intervalu spolehlivosti $[L_D; L_H]$, je stanovení správné. Exploratorní analýza předurčí volbu, zda k testu správnosti využijeme intervalový odhad aritmetického průměru, uřezaného průměru, re-transformovaného průměru, mediánu nebo pivotové polosumy. Interaktivní statistická analýza s vhodným softwarem umožňuje snadno a jednoznačně vyšetřit správnost analytického výsledku.

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie,
Univerzita Pardubice
milan.meloun@upce.cz

LITERATURA

[1] MELOUN, M. – MILITKÝ, J.: Statistické zpracování experimentálních dat. Praha, Plus 1994 (1. Vyd.), East Publishing 1996 (2. vyd.).

[2] MELOUN, M. – MILITKÝ, J.: Kompendium statistického zpracování dat. Praha, Academia 2002.

[3] ADSTAT, TriloByte Statistical Software, s. r. o. Pardubice 1990.

POUŽITÉ VÝRAZY

Normální (Gaussovo) rozdělení
– termín zavedl K. Pearson. N-rozměrné normální rozdělení s parametry μ a Σ (μ je n-rozměrný sloupcový vektor o souřadnicích $\mu_1, \mu_2, \ldots, \mu_n$ a $\Sigma = || \sigma_{kj} ||$ je kladně definitivní matice typu (n, n), je spojité rozložení s charakteristickou funkcí

$$\varphi(t_1, t_2, \ldots, t_n) = \exp\left(i \sum_{j=1}^{n} \mu_j t_j - \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} t_j t_k \sigma_{jk} \right)$$

a hustotou

$$f(x_1, x_2, \ldots, x_n) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} (x_j - \mu_j)(x_k - \mu_k) \sigma^{jk} \right],$$

kde Σ je determinant matice Σ a σ^{jk} jsou prvky matice inverzní k matici Σ, $|| \sigma_{kj} || = \Sigma^{-1}$; n-rozměrné normální rozdělení s parametry μ a Σ se obvykle označuje symbolem $\mathbf{N}_n$ (μ, Σ). Parametr μ je vektorem středních hodnot a matice Σ je rovna kovariační matici rozložení

Laplaceovo rozdělení
– (oboustranné exponenciální rozdělení) vyskytuje se v případech, kdy jsou náhodné veličiny měřeny za podmínek kolísání rozptylu kolem určité střední hodnoty

$$f(x) = \frac{1}{2b} e^{\left(-\frac{|x-a|}{b} \right)}$$

Hustota pravděpodobnosti spojité náhodné veličiny x leží v intervalu $(-\infty, \infty)$ Rozdělení je symetrické podle bodu $x = a$. Střední hodnota Laplaceova rozdělení je $E(x) = a$, rozptyl $D(x) = 2b^2$. Ve srovnání s normálním rozdělením je Laplaceovo rozdělení špičatější a má delší konce

Normalita dat
Zdrojová matice, tj. matice výchozích dat, obsahuje *proměnné* v m sloupcích a *objekty* v n řádcích, na nichž jsou tyto proměnné (vlastnosti) měřeny. Protože měřené proměnné mají různé jednotky, a často se řádově liší, bývá zdrojová matice před zpracováním ještě upravována, *škálována*, a to buď (a) *centrováním*, kdy se od prvků sloupce odečte jejich sloupcový aritmetický průměr, nebo (b) *standardizací* čili *normováním*, kdy se prvky centrovaných sloupců ještě dělí svou sloupcovou směrodatnou odchylkou

Nezávislost dat
– prvky analyzovaného výběru nejsou spojeny žádným skrytým vztahem a byly získány nezávisle na sobě

Homogenita dat
– výběr, jehož všechny prvky pocházejí ze stejného rozdělení s konstantním rozptylem. Odlehlá měření silně zkreslují odhady polohy a zejména rozptylu s^2, takže zcela znehodnocují další statistickou analýzu

Četnost s jakou nastává náhodný jev *A* pro libovolně dlouhou posloupnost pozorování, můžeme charakterizovat podílem r/n, kde n je délka posloupnosti (rozsah výběru) a r je počet např. meteoritů. Číslo r nazýváme *absolutní četnost* a podíl r/n *relativní četnost* výskytu náhodného jevu A ve výběru o rozsahu n. Se vzrůstajícím rozsahem výběru, relativní četnosti se ustalují v blízkosti hodnoty 0,5

Rozptyl, směrodatná odchylka a variační koeficient
Pokud jsou pozorování soustředěna kolem svého průměru, je jejich variabilita malá. Pokud jsou naopak roztroušena ve značné vzdálenosti od průměru, pak je jejich variabilita velká. *Rozptyl* s^2 je průměr čtverců odchylek od průměru. Když však počítáme výběrový rozptyl, nedělíme většinou součet čtverců odchylek výrazem n, ale $(n-1)$, protože tím docílíme lepšího odhadu celkového rozptylu populace (σ^2) Dělitel $(n-1)$ se nazývá *počet stupňů volnosti* rozptylu. Obecný vzorec:

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum x_i \right)^2}{n} \right]$$

Medián – jsou-li pozorování uspořádána vzestupně nebo sestupně, potom *medián* τ je hodnota, která rozdělí pozorování na dvě stejně velké skupiny

Pořádkové statistiky
– prvky výběru z dat, uspořádané vzestupně $x_1 \; x_2 \; \ldots \; x_n$. Platí, že střední hodnota i-té pořádkové statistiky $E(x_{(i)})$ 100 P_i procentnímu kvantilu výběrového rozdělení $Q(P_i)$ a symbol $P_i \; i/(n+1)$ označuje *pořadovou pravděpodobnost.* Připomeňme že 100 P_i *procentní výběrový kvantil* je hodnota pod kterou leží 100 P_i procent prvků výběru

Kvantilově-kvantilový graf
– (graf Q–Q), (osa x: $Q_T(P_i)$, osa y: x_i). Umožňuje posoudit shodu výběrových rozdělení, jež je charakterizováno kvantilovou funkcí $Q_E(P_i)$, s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$. Jako odhad kvantilové funkce výběru se využívají pořádkové statistiky $x_{(i)}$

Rankitový graf
(osa x: kvantil normovaného Gausova rozdělení u_p, osa y: x_i). Pro porovnání rozdělení výběru s rozdělením normálním se Q_Q graf nazývá *grafem rankitovým*

Doporučené reference: www.statsoft.cz, http://ssysel.hyperlink.cz/vyuka/ucebnice/index.htm