Využití exploratorní analýzy v analýze jednorozměrných dat

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice

Souhrn: *Účelem průzkumové analýzy dat (EDA), jako prvního kroku v analýze jednorozměrných dat, je odhalení jejich statistických zvláštností pomocí 12 grafických diagnostik, ověření základních předpokladů a případná mocninná transformace. Při průzkumové analýze složitějších, nákladných nebo unikátních měření je totiž nutné posoudit zvláštnosti chování dat ještě před vlastní, rutinní statistickou analýzou. Jedině tak lze zabránit provádění numerických výpočtů bez hlubších statistických souvislostí. EDA je pak obzvláště důležitá ve stopové analýze mikroelelemntů a kontrole kvality.*

1. Postup analýzy dat ve stopové analýze

V prvním kroku se v průzkumové (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 rozdělení Gaussova. Interaktivní statistická analýza na počítači tento postup ulehčuje, většina statistického software totiž nabízí řadu diagnostických grafů a diagramů.

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 korek-tnějšího 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.

Postup statistické analýzy jednorozměrných dat lze shrnout do čtyř kroků:

A. Průzkumová (exploratorní) analýza dat (EDA)

1. Zkoumání zvláštností dat:
Odhalení stupně symetrie a špičatosti rozdělení;
Indikace lokální koncentrace dat a rozdělení výběru;
Nalezení vybočujících a podezřelých prvků ve výběru.

2. Ověření předpokladů o datech:
Ověření normality rozdělení; Ověření nezávislosti prvků výběru;
Ověření homogenity rozdělení výběru; Určení minimálního rozsahu výběru.

3. Transformace dat:
Mocninná transformace; Box-Coxova transformace.

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

1. Klasické odhady (bodové a intervalové): - momentové;
2. Robustní odhady (bodové a intervalové): - kvantilové, - uřezané, - winsorizované.

A. Postup analýzy rutinních dat

Při zpracování rutinních výsledků měření předpokládáme, že známe rozdělení souboru dat. Předpokládá se, že rozdělení dat je normální a data asi splňují předpoklady nezávislosti a homogenity. Účelem je

a) Test nezávislosti: testování nezávislosti prvků výběru - autokorelace,

b) Test homogenity: testování homogenity výběru,

c) Test normality: 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 *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. 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 byl nalezen optimální exponent λ, provádí se další analýza v této transformaci a nakonec se vyčíslí zpětná transformace do původních proměnných. Nebyla-li mocninná transformace úspěšná, je možné pomocí technik průzkumové analýzy dat nalézt vhodné teoretické aproximující rozdělení a realizovat další postup na základě obecných vztahů pro rozptyl, resp. střední hodnotu. 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.

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.

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.

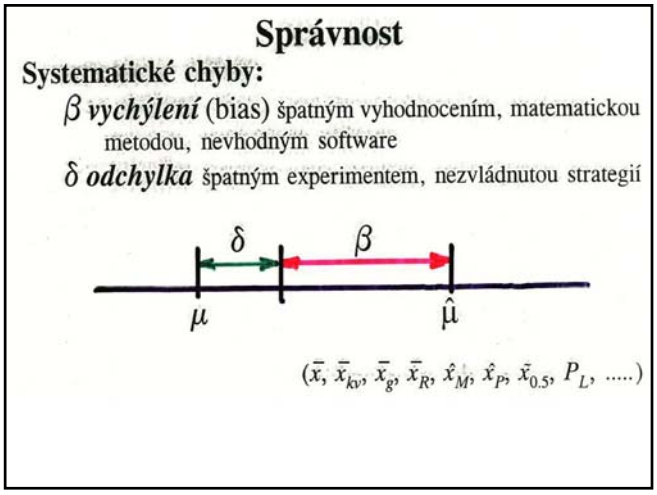
[4] Meloun M., Militký J.: *Statistická analýza experimentálních dat*, Academia Praha 2004, ISBN 80-200-1254-0.

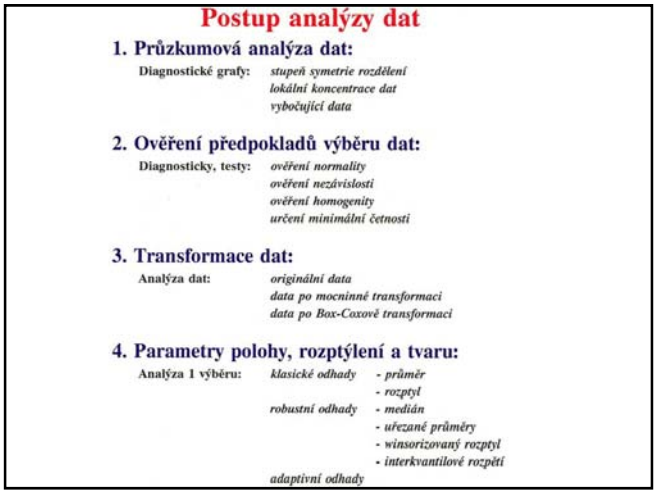
[5] Meloun M., Militký J.: *Kompendium statistického zpracování dat*, Academia Praha 2006, ISBN 80-200-1396-2.

[6] Meloun M., Militký J., Hill M: *Počítačová analýza vícerozměrných dat v příkladech*, Academia Praha 2005, ISBN 80-200-1335-0.

Slidy přednášky

1. Exploratorní průzkumová analýza dat –grafické diagnostiky





PRŮZKUMOVÁ ANALÝZA DAT (EDA)

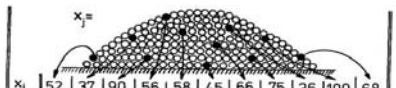
Cílem **průzkumové analýzy dat** je **nalezení zvláštností statistického chování dat** a **ověření jejích předpokladů pro následné zpracování „klasickými" statistickými metodami.**

EDA – Exploratory Data Analysis (Tuckey, Chambers)

Statistické zvláštnosti dat	Základní předpoklady o výběru
Nesymetrie (levostranné nebo pravostranné)	Homogenita rozdělení
	Shoda s teoretickým rozdělením (obvykle normálním)
Lokální koncentrace dat (špičatost nebo plochost)	Potřebná velikost výběru
Extrémní data, odlehlé hodnoty	Nezávislost dat

Reprezentativní náhodný výběr je charakterizován předpoklady:

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



x_j	52	37	90	56	58	45	66	75	36	100	68
j	1	2	3	4	5	6	7	8	9	10	11

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

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

Technika pořadí a hloubek: každá pořádková statistika $x_{(i)}$ má své

rostoucí pořadí $R_{P_i} = i$,

klesající pořadí $K_{P_i} = n + 1 - i$,

hloubku H_i, což je menší číslo z obou uvedených pořadí

$H_i = \min(R_{P_i}, K_{P_i})$.

Pořádkové statistiky: vzestupně setříděné prvky výběru $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$.

Pořadová pravděpodobnost je dána vztahem

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

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

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

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

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

Písmenové hodnoty: kvantily pro pořadové pravděpodobnosti

$$P_i = 2^{-i},\ i = 1, 2, \ldots,$$

i	i-tý kvantil	Pořadová pravděpodobnost P_i	Symbol písmenové hodnoty L
1	medián	$2^{-1} = 1/2$	M
2	kvartily	$2^{-2} = 1/4$	F
3	oktily	$2^{-3} = 1/8$	E
4	sedecily	$2^{-4} = 1/16$	D

H_L celé číslo: dolní kvantil $L_D = x_{(H_L)}$

horní kvantil $L_H = x_{(n+1-H_L)}$

H_L číslo necelé: lineární interpolace podle vztahů

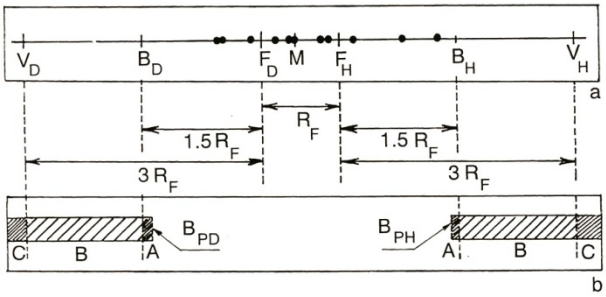
$$L_D = \frac{x_{\mathrm{int}(H_L)} + x_{\mathrm{int}(H_L)+1}}{2}$$

$$L_H = \frac{x_{n+1-\mathrm{int}(H_L)} + x_{n+2-\mathrm{int}(H_L)}}{2}$$

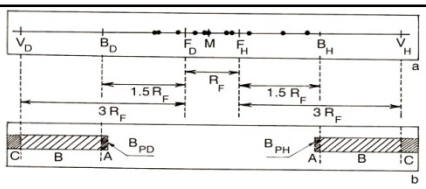
Počet písmenových hodnot n_L: včetně mediánu se určí dle

$$n_L \approx 1.44 \ln(n+1)$$

Užívání písmenových hodnot



Kvartilové rozpětí: $R_F = F_H - F_D$



Přilehlé hodnoty:

$B_H = F_H + 1.5\ R_F$ Horní přilehlá hodnota

$B_D = F_D - 1.5\ R_F$ Dolní přilehlá hodnota

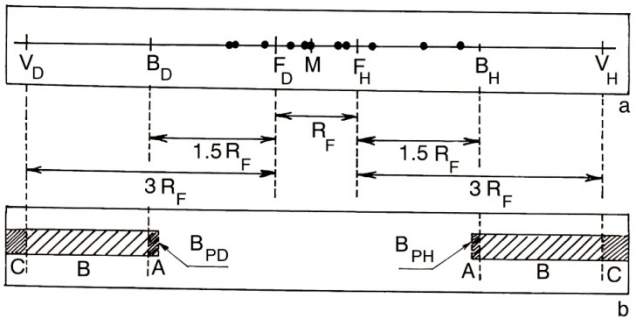
Vnitřní hradby:

$V_H = F_H + 3\ R_F$ Horní vnitřní hradba

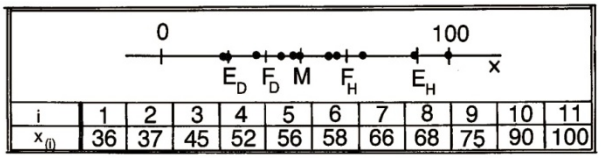
$V_D = F_D - 3\ R_F$ Dolní vnitřní hradba

"Podezřelá měření" (=vybočující hodnoty):

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



Konstrukce barierově-číslicového schématu indikujícího vybočující hodnoty: a) diagram rozptýlení s mediánem M, kvartily F_D (dolní) a F_H (horní) vnitřními hradbami B_D (dolní) a B_H (horní), vnějšími hradbami V_D (dolní) a V_H (horní); b) oblast vybočujících hodnot - A označuje přilehlé body (B_{PD} je blízké B_D a B_{PH} je blízké B_H), B označuje oblast vnějších a C oblast vzdálených bodů.



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

Graficko-tabelární schéma sumarizace dat z předešlé části obrázku:
a) obecné schéma písmenově-číslicového zápisu výběru,
b) sedmipísmenový zápis výběru.

Střední hodnota i-té pořádkové statistiky: rovna $100P_i$ %nímu kvantilu rozdělení výběru $E(x_{(i)}) = F^{-1}(P_i) = Q(P_i)$, kde $F(x)$ je distribuční funkce a $Q(P_i)$ kvantilová funkce výběru.

100α %ní kvantil $\tilde{x}_\alpha$: pro α z intervalu <0, 1> se vyčíslí dle

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

Pro index i musí být splněna nerovnost

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

Identifikace statistických zvláštností výběru dat

(1) **Stupeň symetrie rozdělení výběru**

(2) **Stupeň špičatosti rozdělení výběru**

(3) **Lokální koncentrace dat**

(4) **Přítomnost vybočujících hodnot (měření)**

Pomůcky identifikace statistických zvláštností dat v EDA

Grafické diagnostiky:

Spojité rozdělení:
- G1 Kvantilový graf
- G2 Diagram rozptýlení
- G3 Rozmítnutý diagram rozptýlení
- G4 Krabicový graf
- G5 Vrubový krabicový graf
- G6 Graf polosum
- G7 Graf symetrie
- G8 Graf špičatosti
- G9 Diferenční kvantilový graf
- G10 Graf rozptýlení s kvantily
- G11 Odhad hustoty pravděpodobn.
- G12 Histogram (polygon)
- G13 Kvantil-kvantilový Q-Q graf
- G14 Rankitový graf
- G15 Podmíněný rankitový graf
- G16 Pravděpodobnostní P-P graf
- G17 Kruhový graf

Diskrétní rozdělení:
- G18 Graf poměrů frekvencí
- G19 Poissonův graf
- G20 Modifikovaný Poissonův graf

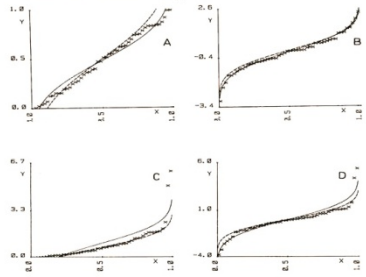
Spojité rozdělení (transformace):
- G21 Hinesové-Hinesův selekční graf
- G22 Graf logaritmu věrohodnost. funkce

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

Grafické diagnostiky EDA

Kvantilový graf (G1)

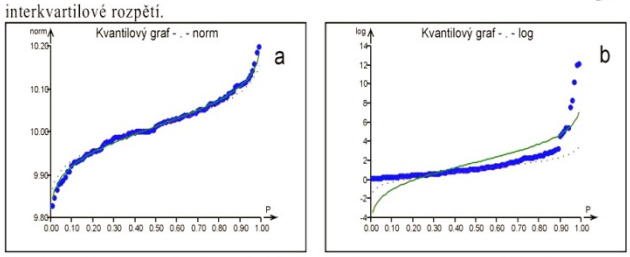
Osa x: pořadová pravděpodobnost P_i,
Osa y: pořádková statistika $x_{(i)}$,



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

Kvantilový graf (osa x: pořadová pravděpodobnost P_i, osa y: pořádková statistika $x_{(i)}$). **(G1)**
Zakreslují se 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 $\mu = \bar{x}$ a $\sigma = s$, 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.

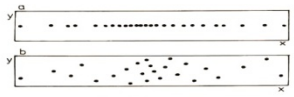


Obr. 2.6 Kvantilový graf (robustní --- a klasický ...) pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Diagram rozptýlení (G2)

Osa x: hodnoty x,
Osa y: libovolná úroveň, obyčejně y = 0

Obsah: jednorozměrná projekce kvantilového grafu do osy x.



Konstrukce a) diagramu rozptýlení a b) rozmítnutého diagramu rozptýlení

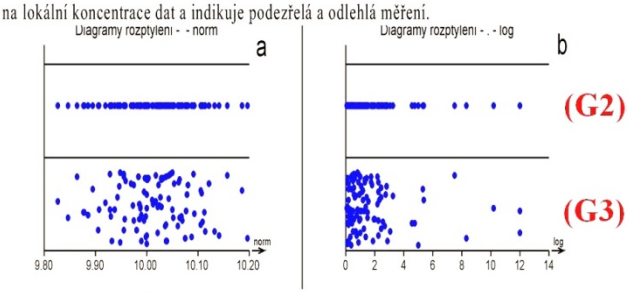
Diagnoza:
(a) přehledně znázornit data, lokální koncentraci dat,
(b) indikuje i podezřelá a vybočující měření.

Rozmítnutý diagram rozptýlení (G3)

Osa x: hodnoty x,
Osa y: libovolný interval náhodných čísel

Obsah: představuje rozmítnutou projekci kvantilového grafu.
Diagnoza: přehledně znázornit data, lokální koncentraci dat, indikuje i podezřelá a vybočující měření.

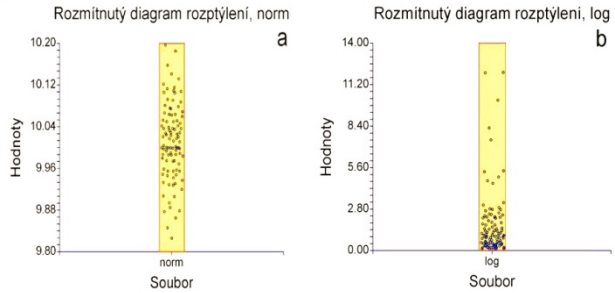
Diagram p rozptýlení (osa x: hodnoty x_i, osa y: libovolná úroveň, např. $y = 0$). Představuje jednorozměrnou rojekci kvantilového grafu do osy x. I při své jednoduchosti ukazuje na lokální koncentrace dat a indikuje podezřelá a odlehlá měření.



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

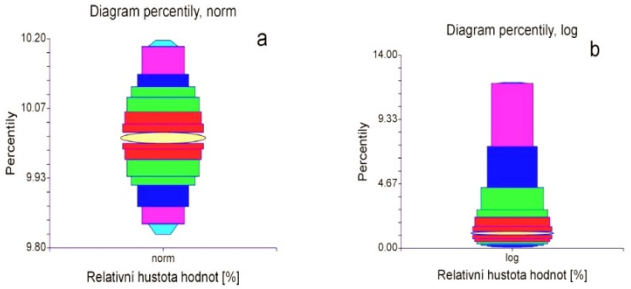
(G3)

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.



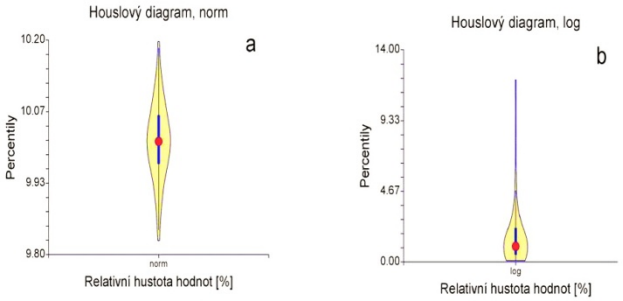
Obr. 2.3 Rozmítnutý diagram rozptýlení pro výběry: (a) *norm*, symetrického (normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000.***

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

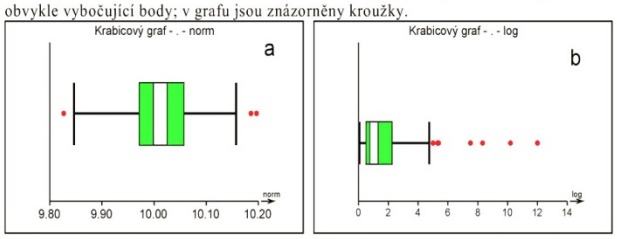


Obr. 2.5 Houslový diagram pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000.***

Krabicový graf (osa x: úměrná hodnotám x, osa y: interval úměrný hodnotě $\sqrt{n}$). **(G4)**
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

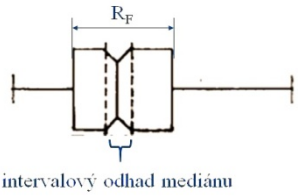
$$B_H = F_H + 1.5\,R_F\,, \quad B_D = F_D - 1.5\,R_F\,.$$

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.***

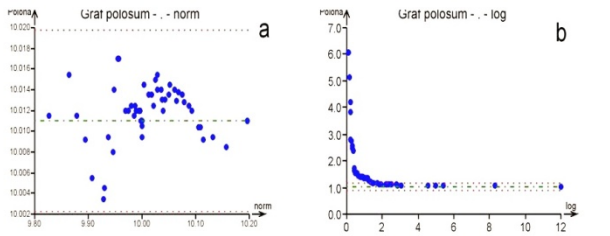
Vrubový krabicový graf (G5)



$$I_{D,H} = M \pm \frac{1{,}57 \cdot R_F}{\sqrt{n}}$$

(G6)

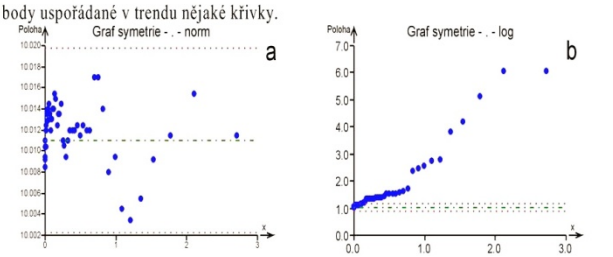
Graf polosum (osa x: pořádkové statistiky $x_{(i)}$, osa y: $Z_i = 0.5\,(x_{(n-1-i)} + x_{(i)})$). Pro symetrické rozdělení je grafem polosum horizontální přímka, určená rovnicí $y = M$. Asymetrické rozdělení vykazuje nenáhodný trend a body pak neoscilují okolo horizontální přímky a měřítko osy y není detailní.



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

(G7)

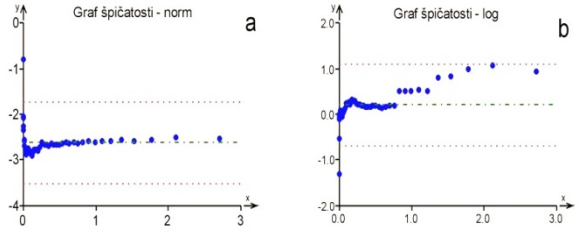
Graf symetrie (osa x: $M - x_{(i)}$, osa y: $x_{(n-1-i)} - M$). Symetrická rozdělení jsou charakterizována přímkou $y = M$. Pro asymetrické rozdělení tato přímka nemá nulovou směrnici a v tomto grafu je směrnice odhadem parametru šikmosti. Asymetrické rozdělení vykazuje body uspořádané v trendu nějaké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



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

Graf špičatosti (osa x: $u_{P_i}^2/2$ pro $P_i = i/(n+1)$, osa y: $\ln(x_{(n+1-i)}/-2u_{P_i})$. **(G8)**

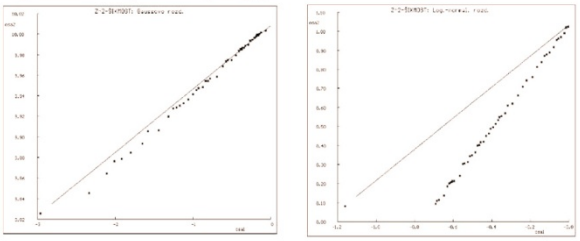
Pro normální rozdělení je grafem horizontální přímka a body leží převážně na této přímce. Pokud body tvoří nenáhodný trend, odpovídá hodnota směrnice této aktuální přímky parametru špičatosti.



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

Graf šikmosti (osa x: $u_{P_i}^2/2$ pro $P_i = i/(n+1)$, osa y: $Z_i = 0.5\,(x_{(n+1-i)} + x_{(i)})$).

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

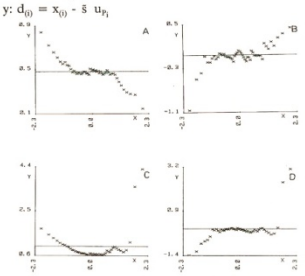


Obr. 2.10 Graf šikmosti pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Diferenční kvantilový graf **(G9)**

Osa x: kvantil u_{P_i},

Osa y: $d_{(i)} = x_{(i)} - \bar{s}\ u_{P_i}$



Diferenční kvantilové grafy pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova, (____ značí normalitu)

Diagnoza:

(a) posouzení rozdělení se špičatostí normálního rozdělení,

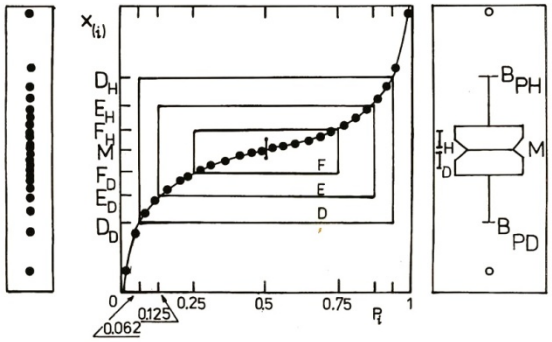
(b) $\bar{s}$ představuje robustní odhad směrodatné odchylky,

(c) horizontální přímka indikuje symetrické rozdělení.

Graf rozptýlení s kvantily **(G10)**

Osa x: P_i,

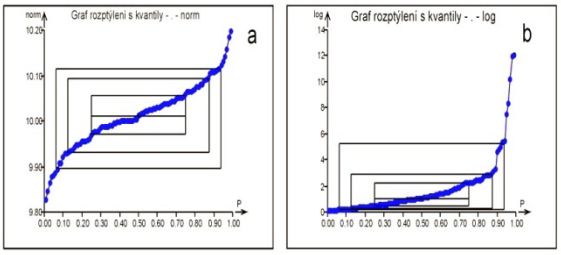
Osa y: $x_{(i)}$



Identifikace řady zvláštností výběru: **(G10)**

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

Graf rozptýlení s kvantily (osa x: P_i, osa y: $x_{(i)}$). Pro symetrická rozdělení má kvantilová funkce sigmoidální tvar. Pro rozdělení zešikmená k vyšším hodnotám je konvexně rostoucí a pro rozdělení zešikmená k nižším hodnotám konkávně rostoucí. **(G10)**



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

Charakteristiky šikmosti a špičatosti

Charakteristiky šikmosti a špičatosti

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 $P_i = 2^{-i}$.

Délka konců (T_E, T_D):

- normální rozdělení (0.534; 0.822),
- rovnoměrné rozdělení (0.405; 0.559)
- Laplaceovo rozdělení (0.693; 1.098).

Šikmost S_L:

- rozdělení sešikmená k vyšším hodnotám S_L záporné,
- rozdělení sešikmená k nižším hodnotám S_L kladné.

Pseudosigma G_L:

- pro rozdělení s delšími konci než normální hodnota G_L roste.

Identifikace rozdělení výběru

Posouzení zvláštností rozdělení výběru:

(1) **Jádrový odhad hustoty pravděpodobnosti**

(2) **Histogram (resp. polygon, rootogram, atd.)**

Porovnání rozdělení výběru s teoretickým rozdělením:

(1) **Kvantil-kvantilový (Q-Q) graf**

(2) **Rankitový graf**

(3) **Podmíněný rankitový graf**

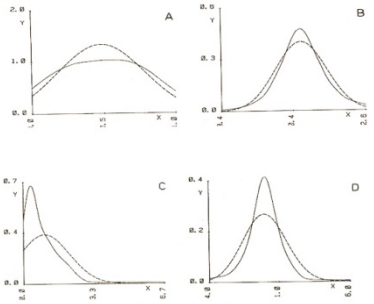
(4) **Pravděpodobnostní graf**

(5) **Kruhový graf**

Jádrový odhad hustoty pravděpodobnosti (G11)

Osa x: proměnná x,

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



Jádrové odhady hustoty pravděpodobnosti pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova.

Čárkovaně je znázorněna hustota Gaussova rozdělení s parametry μ a s^2 a plnou čarou jádrový odhad hustoty pravděpodobnosti empirického rozdělení výběru

(G11)

Pro malé výběry se konstruují jádrové odhady hustoty dle

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

kde šířka pásu h určuje stupeň vyhlazení.

Jádrová funkce K(x) je symetrická kolem nuly.

Bikvadratické jádro se vyčíslí dle

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

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

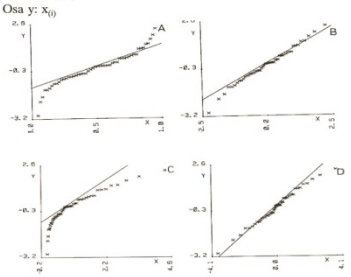
O kvalitě odhadu $\hat{f}(x)$ rozhoduje volba parametru h

$$h_{opt} = 2.34\,\sigma\,n^{-0.2}$$

Kvantil-kvantilový graf (Graf Q-Q) (G13)

Osa x: $Q_S(P_i)$,

Osa y: $x_{(i)}$



Grafy Q-Q pro porovnání rozdělení výběru normálního rozdělení s teoretickým rovnoměrným (A), normálním (B), exponenciálním (C) a Laplaceovým (D) rozdělením. Čarou je vyznačen teoretický průběh

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

kde P_i je pořadová pravděpodobnost.

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

(G13)

Obvykle lze normovat použitím proměnné

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

kde Q je ***parametr polohy***,

R je ***parametr rozptýlení*** mající význam ***měřítka***.

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

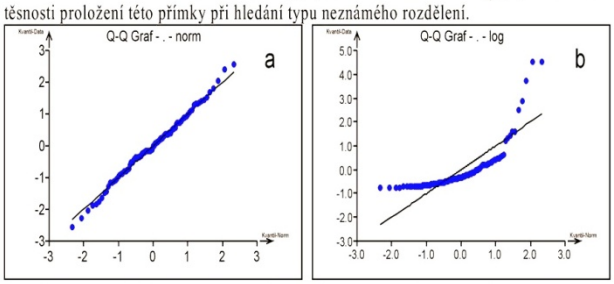
$$x_{(i)} = Q + R\,.\,Q_s(P_i)$$

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

Rozdělení	$F_T(s)$	$f_T(s)$	y	x
rovnoměrné	s	1	$x_{(i)}$	P_i
exponenciální				
x ≥ Q	$1 - \exp(-s)$	$\exp(-a)$	$x_{(i)}$	$-\ln(1 - P_i)$
normální[1]	$\Phi(s)$	$\frac{\exp(-0.5\,s^2)}{\sqrt{2\pi}}$	$x_{(i)}$	$\Phi^{-1}(P_i)$
Laplaceovo				
x ≤ Q	$0.5\,\exp(s)$	$0.5\,\exp(s)$	$x_{(i)}$	$\ln(2\,P_i)$ pro $P_i \le 0.5$
x > Q	$0.5\,(2 - \exp(-s))$	$0.5\,\exp(-s)$	$x_{(i)}$	$-\ln(2\,(1 - P_i))$ pro $P_i > 0.5$
lognormální	$\Phi[\ln(s)]$	$\frac{\exp[-0.5\,\ln(s)^2]}{\sqrt{2\pi}}$	$x_{(i)}$	$\exp[\Phi^{-1}(P_i)]$

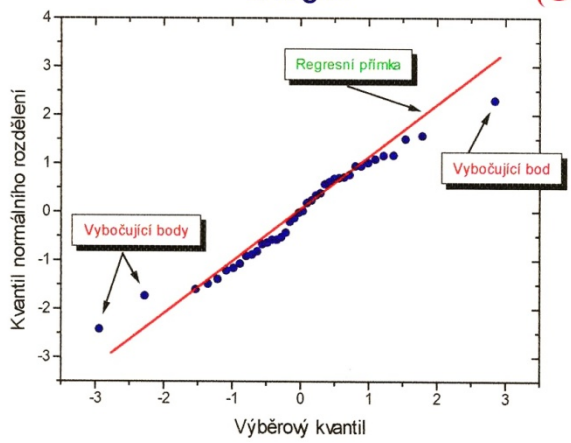
(G13)

Kvantilově-kvantilový graf (graf *Q-Q*) (osa *x*: $Q_T(P_i)$, osa *y*: $x_{(i)}$). Umožňuje posoudit shodu výběrového rozdělení, jež je charakterizováno kvantilovou funkcí $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$. Korelační koeficient r_{xy} je pak kritériem těsnosti proložení této přímky při hledání typu neznámého rozdělení.

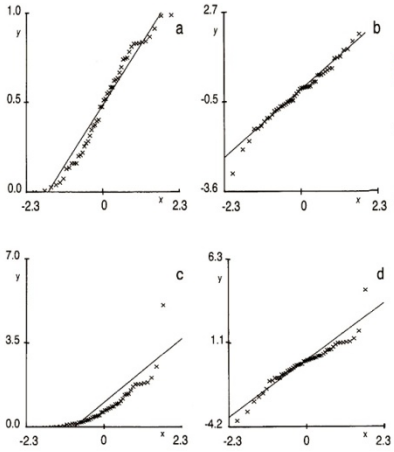


Obr. 2.15 Rankitový čili kvantil-kvantilový graf (*Q-Q* graf) pro ověření shody s teoretickým normálním rozdělením: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT***.

Q-Q graf (G14)



(G14)



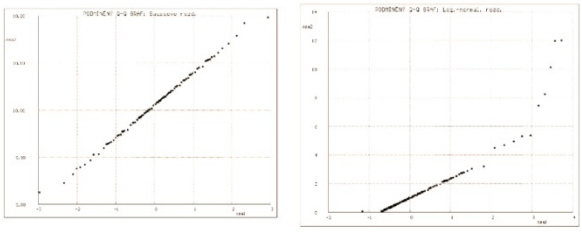
Rankitové grafy pro výběry z rozdělení:

a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova.

Plnou čarou je vyznačen teoretický průběh.

(G15)

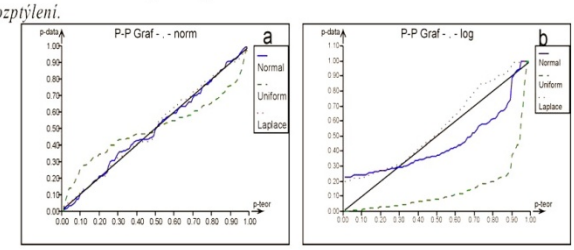
Podmíněný rankitový graf (osa x: Φ^{-1} $[0.5\ (U_{(i-1)} + U_{(i+1)})]$, osa y: $x_{(i)}$). Přibližná lineární závislost je v podmíněném rankitovém grafu důkazem normality testovaného rozdělení výběru. Z grafu normálního rozdělení je patrná výrazně menší lokální variabilita ve srovnání s rankitovými grafy.



Obr. 2.16 Podmíněný rankitový graf pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

(G16)

Pravděpodobnostní graf (***P-P* graf**), (osa x: P_i, osa y: $F_T(S_{(i)})$. 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.*



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

Kruhový graf (G17)

Diagnoza: slouží k ověření, že výběr pochází z rozdělení $F_e(x)$, Transformací

$$Z_{(i)} = F_e(x_{(i)})$$

vyjdou náhodné veličiny $Z_{(i)}$ rozdělené přibližně rovnoměrně na intervalu [0, 1].

Konstrukce: soustava vektorů $\bar{V}_i$ o stejné délce

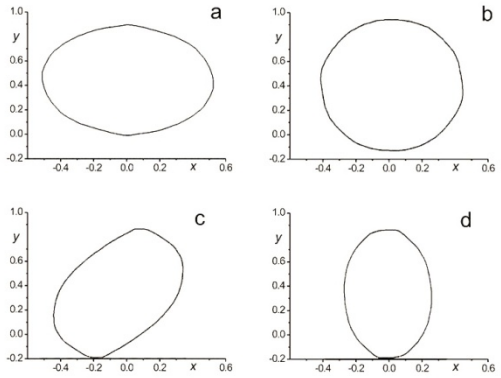
$$l_0 = \left[\frac{N\ (N - 1)}{2}\right]^{-1/2}$$

a směru $\pi\ Z_{(i)}$.
x-ovou a y-ovou složka vektoru $\bar{V}_i$:

$$V_{x_i} = l_0 \cos(\pi\ Z_{(i)}) \qquad V_{y_i} = l_0 \sin(\pi\ Z_{(i)})$$

Úhly se uvažují v radiánech.

(G17)



Kruhový graf pro výběry z rozdělení a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova.

2. Předpoklady o výběru

Ověření předpokladů o datech

Určení minimální velikosti výběru

Rozsah výběru n ovlivňuje přesnost odhadů parametrů polohy

Rozptyly odhadů jsou funkcí n^{-1}.

Metoda volby šíře intervalu přesnosti d:

(1) Z n_1 předběžných hodnot určí odhad rozptylu $s_0^2(x)$.

(2) Zvolí se číslo d tak, aby s pravděpodobností $(1 - \alpha)$ platilo

$$\mu - d \le \bar{x} \le \mu + d.$$

(3) Minimální velikost výběru se vyčíslí dle

$$n_{min} = \left[\frac{t_{1-\alpha/2}(n_1 - 1)}{d}\right]^2 s_0^2(x)$$

kde $t_{1-\alpha/2}(n_1 - 1)$ je kvantil Studentova rozdělení s $(n_1 - 1)$ stupni volnosti.

Ověření nezávislosti prvků výběru

Závislost měření je způsobena:

a) nestabilitou měřicího zařízení,

b) nekonstantností podmínek měření,

c) zanedbáním faktorů: objem vzorků, teplota, nečistota, ...

d) nesprávným, nenáhodným výběrem vzorků k měření,

e) časová závislost mezi prvky výběru.

Test významnosti autokorelačního koeficientu ρ_a:

Hypotéza: nulová H_0: $\rho_a = 0$, a alternativní H_A: $\rho_a \neq 0$.

Testační kritérium:

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

a T je von Neumannův poměr

$$T = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Testování: je-li

$$|t_n| > t_{1-\alpha/2}(n+1)$$

je nutno hypotézu o nezávislosti prvků výběru na hladině významnosti α zamítnout.

ZÁVISLOST A AUTOKORELACE

Obecná definice závislosti:

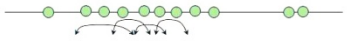
$$x_i = kF(x_1, x_2, \ldots, x_{i-1}) + e_i$$

pokud platí k = 0, jedná se o data nezávislá

Lineární závislost prvků jednoho souboru - **AUTOKORELACE**

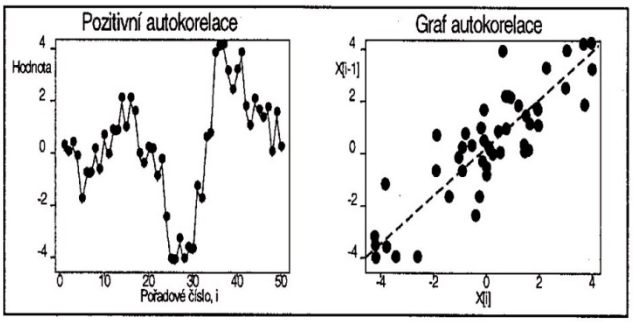
$x_i = \rho_k x_{i-k} + e_i$ $\qquad$ ρ_k autokorelační koeficient k-tého řádu

autokorelace I. řádu sousední hodnoty



autokorelace II. řádu hodnoty „přes jednu"

PŘÍKLADY AUTOKORELACE



Ověření normality výběru

Existují dva základní typy testů normality:

1. Je-li typ odchylek od normality předem znám, používají se tzv. ***směrové testy***.
2. Není-li typ odchylek od normality předem znám, používají se tzv. ***omnibus-testy***.

1. pravidlo: ***Testy jsou obecně vždy méně citlivé na odchylky od normality než diagnostické grafy.***

2. pravidlo: ***Když není normalita rozdělení prokázána, je nutné hlouběji analyzovat data.***

1. Test kombinace výběrové šikmosti a špičatosti.

Hypotéza: nulová H_0: normalita rozdělení výběru, *vs.* H_A: ...

Testovací kritérium: je definováno

$$C_1 = \frac{\hat{g}_1^2}{D(\hat{g}_1)} + \frac{[\hat{g}_2 - E(\hat{g}_2)]^2}{D(\hat{g}_2)}$$

kde výběrová šikmost a její rozptyl $\hat{g}_1$, $D(\hat{g}_1)$, resp. výběrová špičatost a její střední hodnota resp. rozptyl $\hat{g}_2$, $E(\hat{g}_2)$, $D(\hat{g}_2)$.

Testování: při $C_1 > \chi^2_{1-\alpha}(2)$, je nutno hypotézu o normalitě rozdělení výběru zamítnout.

Ověření homogenity výběru

Modifikace vnitřních hradeb B_D^* a B_H^*

$$B_D^* = \tilde{x}_{0.25} - K(\tilde{x}_{0.75} - \tilde{x}_{0.25})$$

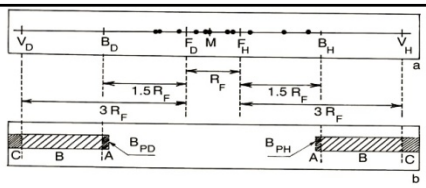
$$B_H^* = \tilde{x}_{0.75} + K(\tilde{x}_{0.75} - \tilde{x}_{0.25})$$

Parametr K: volí se tak, aby pravděpodobnost P(n, K), že z výběru velikosti n 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 $8 \leq n \leq 100$ použít aproximace

$$K \approx 2.25 - \frac{3.6}{n}$$

Pro takto určený parametr K se všechny prvky výběru ležící mimo hradby $[B_D^*, B_H^*]$ považují za vybočující.



Přilehlé hodnoty:

$B_H = F_H + 1.5\ R_F$ Horní přilehlá hodnota

$B_D = F_D - 1.5\ R_F$ Dolní přilehlá hodnota

Vnitřní hradby:

$V_H = F_H + 3\ R_F$ Horní vnitřní hradba

$V_D = F_D - 3\ R_F$ Dolní vnitřní hradba

"Podezřelá měření" (=vybočující hodnoty):

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

3. Transformace dat

Transformace dat

Transformace dat: vede ke stabilizaci rozptylu,
zesymetričtění rozdělení,
k normalitě,

1. ***Stabilizace rozptylu:*** rozptyl $\sigma^2(y)$ lze určit

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

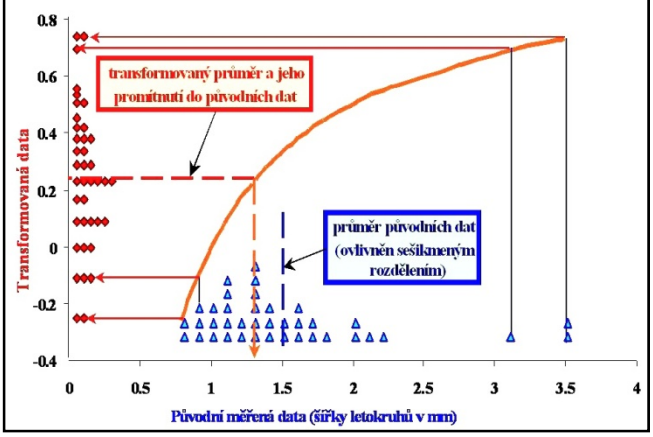
kde C je konstanta. Hledaná transformace g(x) je řešením diferenciální rovnice

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

Instrumentální metody, přístroje: konstantní relativní chyba $\delta(x)$, $\sigma^2(x)$ je dán funkcí $f_1(x) = \delta^2(x)\, x^2 = \text{konst}\ x^2$ a dosazením do $g(x) = \ln x$

2. ***Zesymetričtění rozdělení*** výběru je možné provést jednoduchou mocninnou transformací..

TRANSFORMACE DAT - ukázka



Symetrizující mocninná tranformace

zesymetričtění rozdělení výběru provede

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

1. Nezachovává měřítko,
2. Není k λ všude spojitá,
3. Zachovává pořadí dat ve výběru,
4. Hodí se pouze pro kladná data.

Normalizační Box-Coxova transformace

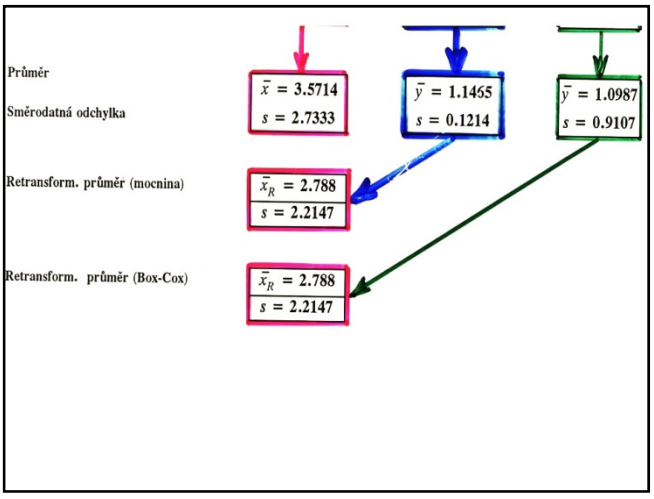
přiblížení rozdělení výběru k normálnímu

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

Transformace
1. Je vzhledem k λ spojitá,
2. Je použitelná pouze pro kladná data, jinak $(x - x_0)$.

Schema mocninné a Box-Coxovy transformace

x	y (mocninná)	y (Box-Cox)
0.5	0.9117	-0.6621
0.9	0.9861	-1.0462
1.0	1.0000	0.0000
1.0	1.0000	0.0000
1.1	1.0128	0.0959
1.3	1.0356	0.2670
1.5	1.0556	0.4166
1.5	1.0556	0.4166
1.5	1.0556	0.4166
1.8	1.0815	0.6114
1.9	1.0893	0.6701
2.0	1.0968	0.7262
2.0	1.0968	0.7262
2.7	1.1416	1.0620
3.2	1.1678	1.2582
3.2	1.1678	1.2582
3.3	1.1726	1.2942
3.3	1.1726	1.2942
3.6	1.1862	1.3968
5.2	1.2459	1.8439
5.5	1.2552	1.9140
5.5	1.2552	1.9140
6.0	1.2699	2.0239
6.0	1.2699	2.0239
7.0	1.2962	2.2217
8.0	1.3195	2.3963
8.0	1.3195	2.3963
11.5	1.3849	2.8869



Zpětná transformace

$$\bar{y},\ s^2(y),\ \bar{y} \pm t_{1-\alpha/2}(n-1).\ s(y)/\sqrt{n}$$

1. **Nekorektní přístup**:

$$\bar{x}_R = \bar{x}_\lambda = \left[\frac{\sum_{i=1}^{n} x_i^{\lambda}}{n}\right]^{1/\lambda}$$

$\bar{x}_R = \bar{x}_{-1}$ představuje *harmonický průměr*,
$\bar{x}_R = \bar{x}_0$ představuje *geometrický průměr*,
$\bar{x}_R = \bar{x}_1$ představuje *aritmetický průměr*,
$\bar{x}_R = \bar{x}_2$ představuje *kvadratický průměr*.

2. **Korektní přístup**: z Taylorova rozvoje funkce $y = g(x)$ v okolí $\bar{y}$,

$$\bar{x}_R \approx g^{-1}\left[\bar{y} - \frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)\right]$$

Pro rozptyl vyjde $s^2(x_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$.

(a) Pro speciální případ $\lambda = 0$, tzn. logaritmickou transformaci typu $g(x) = \ln x$, bude

$$\bar{x}_R \approx \exp\left[\bar{y} + 0.5\ s^2(y)\right]$$

a rozptyl $s^2(x_R) \approx \bar{x}_R^2\ s^2(y)$.

(b) Pro případ $\lambda \neq 0$:

$$\bar{x}_{R,1,2} = \left[0.5(1+\lambda\bar{y}) \pm 0.5\sqrt{1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2(\bar{y}^2 - 2\,s^2(y))}\right]^{1/\lambda}$$

a rozptyl $s^2(x) = \bar{x}_R^{-2\lambda+2}\ s^2(y)$.

Rozptyl se vyčíslí

$$s^2(x_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$$

kde derivace jsou vyčísleny v bodě $x = \bar{x}_R$.

100(1 - α)%ní interval spolehlivosti se vyčíslí dle

$$\bar{x}_R - I_D \le \mu \le \bar{x}_R + I_H$$

kde

$$I_D = g^{-1}\left(\bar{y} + G - t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right)$$

$$I_H = g^{-1}\left(\bar{y} + G + t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right)$$

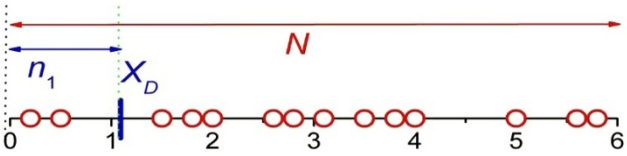
$$G = -0.5\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$$

kde $t_{1-\alpha/2}(n-1)$ je $100(1-\alpha/2)$%ní kvantil Studentova rozdělení s (n-1) stupni volnosti.

4. Analýza jednorozměrných dat

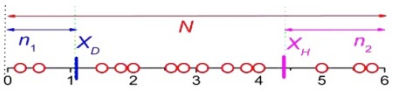
ZÁKLADNÍ ANALÝZA JEDNOROZMĚRNÝCH DAT S OHLEDEM NA LIMITU DETEKCE

MILAN MELOUN,
Katedra analytické chemie, Univerzita Pardubice,
532 10 Pardubice



Ve stopové analýze se požaduje **odhad střední hodnoty** a jeho **interval spolehlivosti** u analytu.

Některé měřené hodnoty jsou pod limitou detekce a existuje pouze kvalitativní informace o koncentraci analytu.



Data mají zvláštnosti:
1) Obsahují často extrémně velké hodnoty, které nemusí být důsledkem chyb měření.
2) Mohou být cenzurována zhora s ohledem na omezený rozsah přístrojů.
3) Jsou vždy kladná a výrazně zešikmená k vyšším hodnotám.
4) Počet je omezen díky malému počtu vzorků a složitému způsobu jejich získávání.
5) Jsou často prostorově nebo časově závislá.
6) Je obtížné opakovat vzorkování a měření za stejných podmínek, protože se koncentrace stopových látek může měnit v čase i v prostoru.
7) Je omezeno použití běžných technik určování odhadů parametrů polohy a rozptýlení ale také je omezena identifikace vybočujících měření.
8) Robustní techniky selhávají, protože eliminují extrémy, které zde nejsou chybami ale důsledkem zešikmení rozdělení dat.

Uvedeme zde problémy spojené s velikostí výběru a zešikmením rozdělení dat.

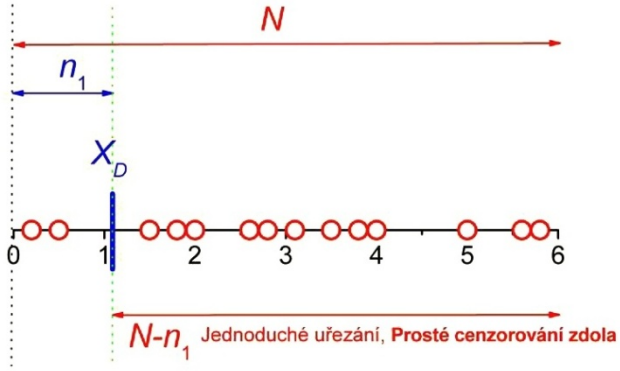
Příklad 1. *Určení koncentrace nečistot v surovině*

Byla sledována koncentrace nečistot v μg/g:

DL, DL, 1.24, 1.49, 1.50, 1.56, 1.61, 1.78,

kde DL značí hodnoty pod limitou detekce x_D = 1 μg/g .

Cíl: odhad střední hodnoty rozptylu a intervalu spolehlivosti za předpokladu normálního rozdělení.



Přehled operací analýzy jednorozměrných dat

A. Průzkumová (exploratorní) analýza dat (EDA):
1. Odhalení stupně symetrie a špičatosti výběrového rozdělení;
2. Indikace lokální koncentrace výběru dat;
3. Nalezení vybočujících a podezřelých prvků ve výběru;
4. Porovnání výběrového rozdělení dat s typickými rozděleními;

B. Ověření předpokladů o datech:
1. Ověření nezávislosti prvků výběru;
2. Ověření homogenity rozdělení výběru;
3. Určení minimálního rozsahu výběru;
4. Ověření normality rozdělení výběru.

C. Transformace dat
1. Mocninná transformace výběru;
2. Boxova-Coxova transformace výběru.

D. Konfirmatorní analýza dat (CDA) - odhady parametrů (polohy, rozptýlení a tvaru)
1. Klasické odhady (bodové a intervalové) z výběru;
2. Robustní odhady (bodové a intervalové) z výběru.
3. Retransformované odhady z výběru.

Diagram rozptýlení

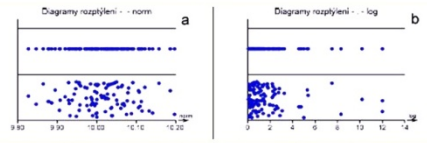
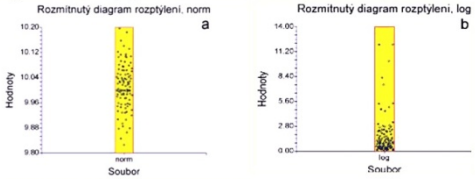


Diagram rozptýlení a rozmítnutý diagram rozptýlení pro výběry (shora dolu): (a) ***norm***, symetrického (normálního), a (b) ***log***, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***



Rozmítnutý diagram rozptýlení pro výběry: (a) ***norm***, a (b) ***log***,

Diagram percentilů

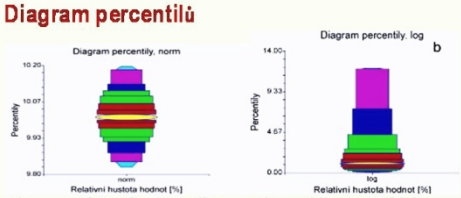
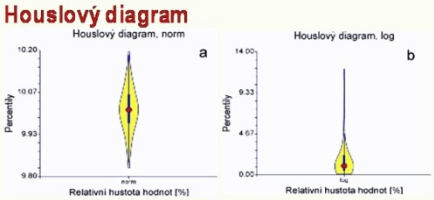


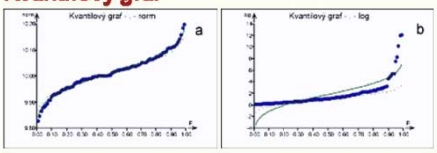
Diagram některých percentilů pro výběry: (a) ***norm***, a (b) ***log***,

Houslový diagram



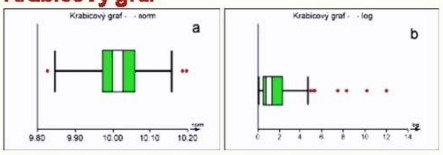
Houslový diagram pro výběry: (a) ***norm***, a (b) ***log***,

Kvantilový graf



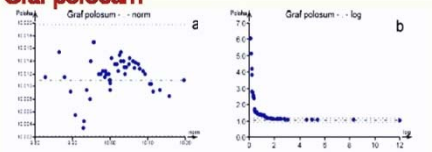
Kvantilový graf (robustní --- a klasický ...) pro výběry: (a) *norm*, a (b) *log*,

Krabicový graf



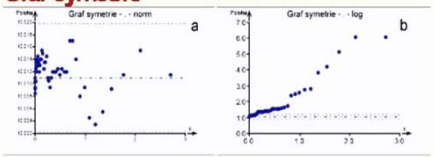
Vrubový krabicový graf pro výběry: (a) *norm*, a (b) *log*,

Graf polosum



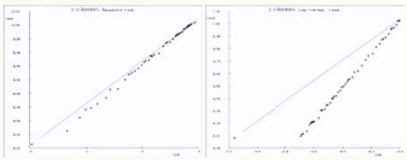
Graf polosum pro výběry: (a) *norm*, a (b) *log*,

Graf symetrie



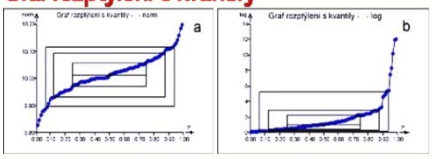
Graf symetrie pro výběry: (a) *norm*, a (b) *log*,

Graf šikmosti



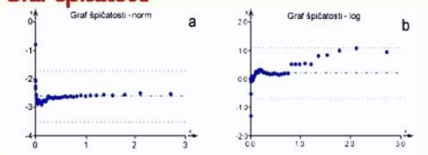
Graf šikmosti pro výběry: (a) *norm*, a (b) *log*,

Graf rozptýlení s kvantily



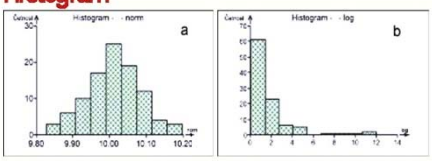
Graf rozptýlení s kvantily pro výběry: (a) *norm*, a (b) *log*,

Graf špičatosti



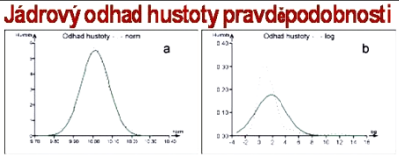
Graf špičatosti pro výběry: (a) *norm*, a(b) *log*,

Histogram



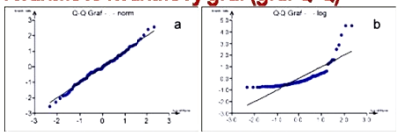
Histogram pro výběry: (a) *norm*, a (b) *log*,

Jádrový odhad hustoty pravděpodobnosti



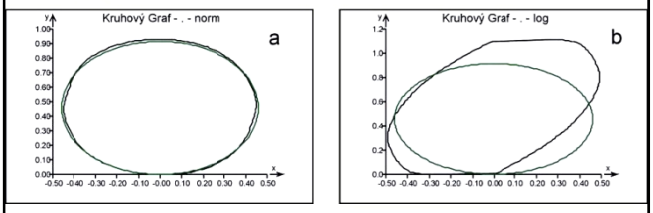
Jádrový odhad hustoty pravděpodobnosti pro výběry s empirickou (čárkovaně) a aproximační křivkou Gaussova rozdělení (plná čára): (a) *norm*, a (b) *log*,

Kvantilově-kvantilový graf (graf *Q-Q*)

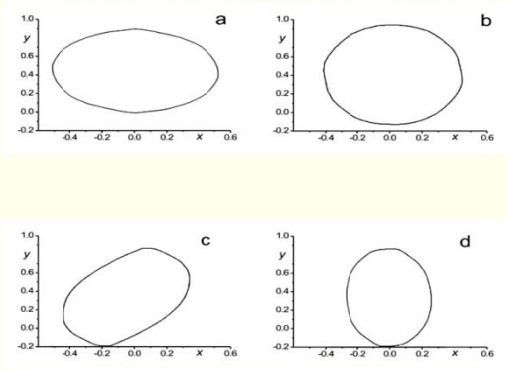


Rankitový čili kvantil-kvantilový graf (*Q-Q* graf) ověření shody s teoretickým normálním rozdělením: (a) *norm*, a (b) *log*,

Kruhový graf



Kruhový graf



Kruhový graf pro výběry z rozdělení (a) rovnoměrného, b) normalního, (c) exponenciálního a (d) Laplaceova.

Náhodný výběr reprezentovaný N-ticí dat $x_1, x_2, \ldots, x_N$.

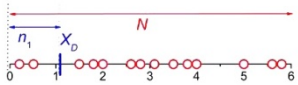
Střední hodnotu, rozptyl a interval spolehlivosti střední hodnoty lze spolehlivě vypočítat jen při znalosti typu rozdělení pravděpodobnostního modelu měření.

Existují dvě mezní situace:

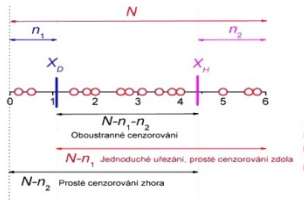
1) Rozmezí dat je v rámci jednoho řádu. Užijeme standardní statistické metody za konstantního rozptylu a aditivního modelu měření.
2) Rozmezí dat je v rozmezí několika řádů. Použijeme logaritmickou transformaci dat nebo multiplikativní model měření.

Někdy nelze získat všechny výsledky měření,

protože některé jsou pod mezí detekce $DL = x_D$.



Náhodný výběr reprezentovaný N-ticí dat $x_1, x_2, \ldots, x_N$

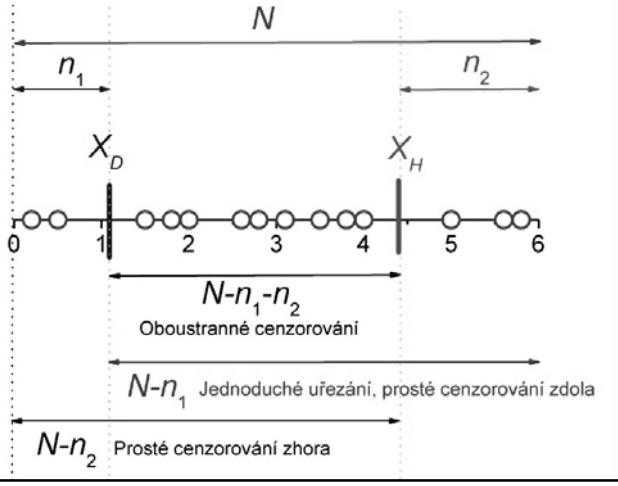


N-n₁ značí:
Cenzorování typu I
Cenzorování typu II

1) Z celkového počtu měření N je $N - n_1$ nad limitou detekce a zbytek n_1 je pod limitou detekce, **prosté cenzorování zdola**.
2) Při omezení shora je maximálně meřitelná hodnota $UL = x_U$ a n_2 měření je nad limitou intervalu stanovení, **prosté cenzorování shora**.
3) U **oboustranného cenzorování** jsou známy hodnoty pouze pro $N - n_1 - n_2$ prvků výběru.

Cohen definuje u prostého cenzorování zdola tři základní úlohy:
1) Data pod limitou detekce n_1 a ani celkový počet N nejsou známa. Je k dispozici pouze $N - n_1$ hodnot, **jednoduché uřezání**.
2) Je znám počet hodnot n_1 pod mezí detekce x_D a dále hodnoty $N - n_1$ prvků výběru celkové velikosti N, **cenzorování typu I.**
3) Nejmenších n_1 prvků pod mezí detekce nemá hodnotu. Pouze je známa hodnota x_D, která je menší než než nejmenší hodnota prvků nad limitou detekce $x_{(n1+1)}$, kde $x_{(n1+1)}$ značí pořádkovou statistiku, **cenzorování typu II**.

Standardně se řeší pouze cenzorování typu I.



Vliv zešikmení dat na intervaly spolehlivosti

Standardní zpracování výběrů: odhad aritmetického průměru x_A a výběrového rozptylu s^2.

1) Výběr velikosti N pochází z **nenormálního rozdělení** se střední hodnotou μ a rozptylem σ^2, má náhodná veličina $Z = \sqrt{N}(\bar{x} - \mu)/\sigma$ asymptoticky normální rozdělení.

2) σ není známo a nahrazuje se výběrovou směrodatnou odchylkou s a Studentova náhodná veličina $t = \sqrt{N}(\bar{x} - \mu)/s$ má Studentovo rozdělení s N-1 stupni volnosti.

Dvě základní **techniky omezení vlivu zešikmení dat**:

1. Použití **vhodné transformace** vedoucí ke zlepšení statistických vlastností testovacích statistik. Postup není prost omezení a vždy je využito rozvoje do řady a použito několika prvních členů.

2. Pro *CI* se používá **klasický interval spolehlivosti** pro korigovaný průměr, který je blíže střední hodnotě, a to vyšší než $\bar{x}$.

100(1 - α)% na interval spolehlivosti *CI* je definován vztahem

$$P(CID \le \mu \le CIH) = 1 - \alpha$$

Symbol P(.) označuje pravděpodobnost a α je hladina významnosti.
Obyčejně se volí $\alpha = 0.05$ nebo $\alpha = 0.01$ s tím, že čím je α menší, tím je interval (*CID, CIH*) širší.

1) **Rozptyl σ^2 je znám** a interval spolehlivosti *CI* se vyjádří ve tvaru

$$\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{N}} \le \mu \le \bar{x} - z_{\alpha/2}\frac{s}{\sqrt{N}}$$

kde $z_{1-\alpha/2} = -z_{\alpha/2}$ jsou kvantily normovaného normálního rozdělení.

2) **Rozptyl σ^2 není znám** a interval spolehlivosti *CI* se vyjádří ve tvaru

$$\bar{x} - t_{1-\alpha/2}(N-1)\frac{s}{\sqrt{N}} \le \mu \le \bar{x} - t_{\alpha/2}(N-1)\frac{s}{\sqrt{N}}$$

kde $t_{1-\alpha/2}(N-1) = -t_{\alpha/2}(N-1)$ jsou kvantily Studentova rozdělení s N-1 stupni volnosti.

Omezení asymetrie rozdělení Studentovy statistiky

Asymetrie rozdělení *t* statistiky je zřejmá z Edgeworthova rozvoje. Johnson navrhl vztah

$$t_J = \sqrt{N}[(\bar{x} - \mu) + \frac{g_1(x)s}{6N} + \frac{g_1(x)}{3s}(\bar{x} - \mu)^2]/s$$

Pro tuto transformaci již přibližně platí, že $P(t_J \le x) \approx F_n(x)$.

Některé problémy eliminují transformace navržené Hallem

$$t_H = K + \frac{g_1(x) * K^2}{3} + \frac{g_1(x)^2 * K^3}{27} + \frac{g_1(x)}{6N},$$

resp. $$t_{H1} = \frac{g_1(x)}{6N} + \frac{3 * \sqrt{N} * exp(\frac{2 * K * g_1(x)}{3\sqrt{N}} - 1)}{2 * g_1(x)}, \quad \text{kde } K = \frac{\bar{x} - \mu}{s}.$$

Obě transformace násobené faktorem $N^{-0.5}$ vedou k přibližné normalitě (redukci šikmosti) a jsou invertovatelné .

Inverzní forma statistiky t_H se zahrnutou násobivou konstantou má tvar

$$t_H^{-1}(y) = \frac{3 * \sqrt{N}}{g_1(x)}[(1 + g_1(x) * (\frac{y}{\sqrt{N}} - \frac{g_1(x)}{6N}))^{1/3} - 1]$$

Při sledování úrovně škodlivin je prakticky zajímavý pouze pravostranný interval spolehlivosti (jednostranný interval spolehlivosti zprava tj. horní hranici střední hodnoty). Tento interval se často používá u rozdělení zešikmených vpravo k určení povolené horní hranice např. znečištění.

Pro horní mez pravostranného intervalu spolehlivosti pak platí, že

$$\mu \le \bar{x} + t_H^{-1}(z_{1-\alpha})\frac{s}{\sqrt{N}}$$

Místo normovaného normálního kvantilu *z* se doporučuje použít odpovídajícího kvantilu určeného z Bootstrap výběrů . Místo transformace lze použít zjednodušenou verzi

$$t_a^{-1}(y) = y - \frac{g_1(x)(y^2/3 + 1/6)}{\sqrt{N}}$$

a opět je možno použít Bootstrap kvantilů. Jak je patrné znalost šikmosti výběrového rozdělení je zde nezbytnou podmínkou pro použití korekce.

Uvedeme vztah mezi nejistotou pokrytí zleva, zprava a také z obou stran:

Nejistota pokrytí zprava *NP* vyjadřuje pravděpodobnost, že skutečná střední hodnota je nižší než meze intervalu spolehlivosti, $NP = \alpha/2 + [-0.73 + 0.71\exp(-\alpha/2)]g_1/\sqrt{N}$.

Nejistotu pokrytí zleva *NL* vyjadřuje pravděpodobnost, že skutečná střední hodnota je vyšší než meze intervalu spolehlivosti, $NL = \alpha/2 + [0.19 + 0.026\ln(\alpha/2)]g_1/\sqrt{N}$.

Nejistota pokrytí z obou stran *NC* je pak sjednocení obou chyb pokrytí, tj. *NC=NP+NL*.

Lze určit: 1) Potřebná velikost výběru, aby byla zachována nejistota pokrytí jako rozdíl mezi požadovanou pravděpodobností pokrytí (např. *0.95*) a dosaženou (např. *0.94*).

2) Fixovat nejistotu pokrytí na zvolené hodnotě a pro známé *N* a $g_1(x)$ nalézt pravděpodobnost α^* pro výpočet kvantilu Studentova rozdělení.

Klasický interval spolehlivosti je $\mu \le \bar{x} + t_{1-\alpha}(N-1)\frac{s}{\sqrt{N}}$.

Po dosazení za *NL* = 0.05 bude $0 = \alpha^* + [0.19 + 0.026\ln(\alpha^*)]g_1(x)/\sqrt{N} - 0.05 = f(\alpha^*)$.
Kořenem funkce $f(\alpha^*)$ je pak α^*, pro kterou vypočteme opravený kvantil Studentova rozdělení, čili $t_{1-\alpha^*}(N-1)$.

Výpočet korigovaného průměru

Korigovaný průměr $\bar{x}_0$ pro stanovení intervalu spolehlivosti u asymetrických rozdělení je založen na Johnsonově transformaci a má tvar $\bar{x}_0 = \bar{x} + \frac{s.g_1}{6N}$. Korekce souvisí se šikmostí g_1 a počtem měření *N*.

Odhady minimalizující penále za přecenění či nedocenění odhadu střední hodnoty:

1) Chenová zavedla **MCE-odhad** $\bar{x}_{MCE}$ ve tvaru $\bar{x}_{MCE} = x_A + d.s$, kde *d* se počítá dle

$$d = 0.5\left[b - \frac{2\sqrt{N}}{g_1(x)} + \sqrt{4 - \frac{b^2}{3} + \frac{4.N}{g_1(x)} + \frac{8.\log(a).\sqrt{N}}{b.g_1(x)}}\right]$$

a volba *a* a *b* se doporučuje *a = 10* a *b = 3*.

2) **Penalizovaný průměr** $\bar{x}_P$ se vypočte dle $\bar{x}_P = \bar{x} + \frac{4.5s^2}{\sqrt{N}}f(\bar{x})[1 - F(\bar{x})]$,
kde $f(\bar{x})$ a $F(\bar{x})$ jsou hustoty pravděpodobnosti a distribuční funkce, které nahradíme neparametrickými odhady $f(x) = \frac{\text{int}(\sqrt{N})}{2.N.A(x)}$, kde $A(\bar{x})$ je *k*-tý nejmenší rozdíl $w_i = abs(x_i - \bar{x})$ a $k = int(N^{0.5})$ je vlastně *k*-tá pořádková statistika. Distribuční funkce $F(\bar{x})$ se počítá jako počet hodnot prvků výběru ležících pod $\bar{x}$ dělený *N*.

3) Použití **upraveného výběru uvažujícího extrémy**, kdy se nejvyšší pořádková statistika $x_{(N)}$ nahrazuje hodnotou $\bar{x}$ +4.5 *s*, pokud je větší. Doporučuje se pro silně zešikmená rozdělení, kde se

Pro normální rozdělení mají intervaly spolehlivosti přesně 100(1-α)% pokrytí střední hodnoty.

Jen 100α/2% případech je střední hodnota **menší než CI (nejistota NP zprava)**,
a v 100α/2% případech je střední hodnota **větší než CI (nejistota NL zleva)**.

1. Pro vliv šikmosti na rozdělení náhodné veličiny Z lze použít prvního členu Edgeworthova rozvoje

$$P(Z \le x) = F_n(x) - \frac{g_1(x) * (x^2 - 1)}{6\sqrt{N}} f_n(x),$$

kde $F_n(x)$ je distribuční funkce normovaného normálního rozdělení a $f_n(x)$ je hustota pravděpodobnosti.

Šikmost náhodné veličiny Z je dána vztahem $g_1(Z) = g_1(x)/\sqrt{N}$ a platí pravidlo
čím je $g_1(Z)$ blíže k nule, tím je rozdělení veličiny Z bližší normálnímu.

Pro rozdělení zešikmené k vyšším hodnotám je $g_1(x)$ kladné,
a také rozdělení veličiny Z je zešikmené k vyšším hodnotám, tj. $g_1(Z)$ kladné.
Interval spolehlivosti má vyšší horní mez *CIH* a vyšší dolní mez *CID* než odpovídá rozdělení statistiky Z.

Například: pro výběr rozsahu N =10 ze standardizovaného exponent. rozdělení o $g_1(x)=2$, je 97.5% kvantil rozdělení veličiny Z = 2.24 a odpovídající kvantil normovaného normálního rozdělení je 1.96. Podobně jeho 2.5% kvantil Z =–1.65 proti kvantilu normovaného normálního rozdělení -1.96.
Interval spolehlivosti je celý posunut doprava oproti skutečnému.

2. Pro vliv šikmosti na rozdělení náhodné veličiny t lze použít prvního členu Edgeworthova rozvoje

$$P(t \le x) = F_n(x) + \frac{g_1(x) * (2x^2 + 1)}{6\sqrt{N}} f_n(x)$$

kde $F_n(x)$ je distribuční funkce normovaného normálního rozdělení a $f_n(x)$ je hustota pravděpodobnosti.

Při porovnání s předešlou rovnicí s veličinou Z je patrné opačné znaménko korekčního členu,
čili pro rozdělení dat zešikmené k vyšším hodnotám je $g_1(x)$ kladné a
rozdělení náhodné veličiny t je zešikmené k nižším hodnotám, tj. $g_1(t)$ záporné.
Interval spolehlivosti s veličinou t pak má nižší horní mez *CIH* a nižší dolní mez *CID* než odpovídá reálnému rozdělení statistiky t.
Interval spolehlivosti je proto celý posunut doleva oproti skutečnému.

Vysvětlení: příčinou rozdílu mezi chováním náhodné veličiny Z a t je korelace mezi odhady $\bar{x}$ a s.
Asymptotický korelační koeficient je $\rho(\bar{x}, s) = \frac{g_1(x)}{\sqrt{(g_2(x) - 1}}$, kde $g_2(x)$ je špičatost rozdělení dat.
Problémy výpočtu intervalů spolehlivosti střední hodnoty nastávají u ne-normálního rozdělení, zešikmeného vpravo a „malých velikostí výběru".

Pojem „malá velikost výběru" závisí na šikmosti rozdělení dat.

Modely měření

1. Aditivní model měření: nejčastější model měření x je $x = \mu + \varepsilon$, kde μ je skutečná hodnota (koncentrace analytu) a ε je náhodná chyba s rozdělením o hustotě pravděpodobnosti $f(\varepsilon)$.
Předpokládá se:

a) Předpoklad o konstantnosti rozptylu $\Delta \approx \sigma$. Pro aditivní model je relativní chyba $\delta = \Delta / x$.

b) Předpoklad o nezávislosti měření je většinou splněn. Potíže mohou nastat u vzniku **autokorelace I.řádu** $\varepsilon_i = \rho * \varepsilon_{i-1} + u_i$, kde u_i je náhodná veličina s konstantním rozptylem.

Autokorelační koeficient ρ se odhaduje $\hat{\rho} = \frac{\sum_j^{N-1} (x_j - \bar{x}) * (x_{j+1} - \bar{x})}{[s^2 (N-1)]}$.

Pravidlo: pokud $\hat{\rho}$ leží v intervalu $-2/\sqrt{M} \le \hat{\rho} \le 2/\sqrt{M}$, lze ρ považovat za nevýznamný.

c) Předpoklad normálního rozdělení chyb je potřebný pro konstrukci intervalů spolehlivosti (nejistota) a testování hypotéz. Častější bývá **asymetrické rozdělení dat zešikmené k vyšším hodnotám**.

Asymetrii odstraníme vhodnou transformací $h(x)$, ale ta vede ke vzniku nekonstantního rozptylu

$D(h(x)) = \left[\frac{dh(x)}{dx}\right]^2 * \sigma^2$. Např. pro transformaci $h(x)=\ln(x)$ bude $D(h(x)) = \left(\frac{\sigma}{x}\right)^2 = \delta^2$.

2. Multiplikativní model měření: založen na konstantní relativní chybě a nezápomosti měření $x = \mu * \exp(\varepsilon)$. Po logaritmické transformaci dostaneme aditivní model v logaritmech

$$\ln(x) = \ln(\mu) + \varepsilon.$$

Nevýhody: pro velmi nízké koncentrace μ vychází absolutní chyba měření příliš nízká.

Pokud platí, že $\ln(x) \approx N(\nu, \tau^2)$ má výsledek měření x lognormální rozdělení

s parametry $\mu = \exp(\nu + \tau^2/2)$ a $\sigma = \mu^2(\exp(\tau^2) - 1)$.

K charakterizaci polohy užijeme také geometrický průměr

$$\mu_G = \exp(E(\ln(x))) = \exp(\nu).$$

Geometrický průměr je roven mediánu

$$med(x) = \exp(\nu) = \mu_G.$$

Pro odhady parametrů se používá dat v logaritmické transformaci

$$\hat{\nu} = \frac{1}{N}\sum_{i=1}^{N} \ln x_i, \quad \hat{\tau} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (\ln x_i - \hat{\nu})^2}, \quad \hat{\mu} = \exp(\hat{\nu} + \hat{\tau}^2/2)$$

Pro rozptyly odhadů platí

$$D(\hat{\mu}) = \frac{\mu^2}{N[\exp(\tau^2) - 1]}, \quad D(\hat{\mu}_G) = E(\hat{\mu}_G)^2 [\exp(\tau^2/N) - 1]$$

Střední hodnota geometrického průměru je $E(\hat{\mu}_G) = \exp(\nu + \tau^2/2N)$.

Oba průměry jsou vychýlené odhady.

Geometrický průměr je pro tento model měření výhodnější než aritmetický průměr.

V případě, že se použije „nevhodná" logaritmická transformace bude

$$\ln(x) = \ln(\mu + \varepsilon) = \ln\mu + \ln(1 + \varepsilon/\mu).$$

S využitím Taylorova rozvoje lze psát

$$\ln(1 + x) = x - x^2/2 + x^3/3 + x^4/4$$

a pak $\ln(x) \approx \ln(\mu + \varepsilon) = \ln\mu + \varepsilon/\mu - 0.5 * (\varepsilon/\mu)^2 + (\varepsilon/\mu)^3/3 - (\varepsilon/\mu)^4/4 \ldots$

Odhady parametrů

Rozdělení měření pro oba modely obsahuje střední hodnotu μ a rozptyl σ^2 a jejich odhady získáme:

1) Momentová metoda spočívá v porovnání výběrových a teoretických momentů rozdělení.
Pro normální rozdělení a aditivní model: odhadem je aritmetický průměr x_A a výběrový rozptyl s^2.

2) Metoda maximální věrohodnosti získává odhad maximalizující logaritmus věrohodnostní funkce L

$$\ln L(\mu, \sigma^2, a) = \sum \ln f(x, \mu, \sigma^2, a).$$

Maximálně věrohodné odhady parametrů jsou určeny z maxima věrohodnostní funkce soustavy m

$$\frac{\partial \ln L(\mu, \sigma^2, a)}{\partial \mu} = 0 \qquad \frac{\partial \ln L(\mu, \sigma^2, a)}{\partial \sigma^2} = 0 \qquad \frac{\partial \ln L(\mu, \sigma^2, a)}{\partial a} = 0$$

kde pro normální rozdělení a nezávislá měření má logaritmus věrohodnostní funkce tvar

$$\ln(L) = (-N/2)[\ln(2\pi) + \ln(\sigma^2)] - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}.$$

Pro první derivace logaritmu věrohodnostní funkce pak platí, že

$$\frac{\partial \ln(L)}{\partial \mu} = 2\sum_i x_i - 2N\mu \quad \text{a} \quad \frac{\partial \ln(L)}{\partial \sigma^2} = \frac{\sum_i (x_i - \mu)^2}{[\sigma^2\sqrt{2}]^2} - \frac{N}{2\sigma^2}$$

maximálně věrohodné odhady střední hodnoty a rozptylu pro normální rozdělení jsou opět totožné s výběrovým průměrem a výběrovým rozptylem.

3) Využití pořádkových statistik $x_{(1)} < x_{(2)} < \ldots < x_{(N)}$, které jsou vlastně vzestupně setříděnými hodnota prvků výběru.

Pokud je $F_e(x)$ neznámá výběrová distribuční funkce, má náhodná veličina

$z_{(i)} = F_e(x_{(i)})$ nezávisle na této distribuční funkci rozdělení Beta Be [i, N-i+1]

a její střední hodnota je $E(z_{(i)}) = i/(N+1)$, kde E(.) je operátor střední hodnoty.

Pořádkové statistiky $z_{(i)}$, $z_{(j)}$, $i, j = 1, \ldots N$, jsou korelované s kovariancí závislou na i, j a N.

Pro pořádkové statistiky platí $E(x_{(i)}) = F_e^{-1}(z_{(i)}) = Q_e(P_i)$,

kde $Q_e(P_i)$ je kvantilová funkce a pořadová pravděpodobnost je rovna $P_i = i/(N+1)$.

Pořádková statistika $x_{(i)}$ je hrubým odhadem kvantilu $Q_e(P_i)$ pro pravděpodobnost P_i čili $x_P = Q_e(P)$.

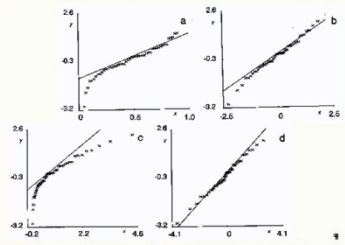
Pokud je třeba získat odhad kvantilové funkce mimo polohy pořádkových statistik pro zadanou pravděpodobnost p, kde $i/(n+1) < P < (i+1)/(n+1)$ volí se **lineární interpolace** dle vztahu

$$x_{(P)} = (N+1)\left(\frac{PN + P - i}{N+1}\right)(x_{(i+1)} - x_{(i)}) + x_{(i)}$$

a její rozptyl $D(x_P)$ se vyčíslí dle vztahu $D(x_P) = \frac{P(1-P)}{N f^2(x_P)}$.

Pořádkové statistiky využijeme pro řešení dvou úloh:
1) stanovení typu rozdělení výběru,
2) odhad jeho vybraných parametrů při využití tzv. **Q-Q grafů.**

Kvantil-kvantilový graf (Q-Q graf): pokud je rozdělení výběru F_e totožné s teoretickým rozdělením F_t, musí platit i shoda kvantilů čili $x_{(i)} = F_t^{-1}(P_i)$ a závislosti bude přímka.



Grafy Q-Q pro porovnání rozdělení výběru normálního rozdělení s (a) teoretickým rovnoměrným, (b) teoretickým normálním, © teoretickým exponenciálním a (d) teoretickým Laplaceovým rozdělením. Čarou je vyznačen teoretický průběh.

1) **Stanovení typu rozdělení**: pro výpočet $F_T^{-1}(P_i)$ je třeba znát obecně parametry teoretického rozdělení.

V řadě případů je však možná standardizace $s = (x - Q)/R$, kde R je parametr rozptýlení.

Standardizované kvantilové funkce $Q_S(P_i) = Fs_T^{-1}(P_i)$ obsahují jen tvarové faktory.

V případě shody obou rozdělení pak resultuje přímková závislost

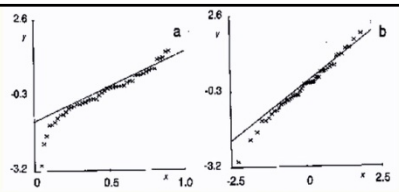
$$x_{(i)} = Q + R \,.\, Q_S(P_i) = a + b \,.\, Q_S(P_i).$$

2) **Odhady parametrů polohy a rozptýlení**: odhad střední hodnoty odpovídá absolutnímu členu a a odhad směrodatné odchylky směrnicí b regresní přímky.

Pro odhad parametrů z Q-Q grafů je možno použít bud MNČ.

Transformací $z = \ln[(1+\sigma * Y)^{1/\sigma}]$ má veličina z normované normální rozdělení N(0,1).

Pořádková statistika $x_{(i)}$ pak souvisí s pořádkovou statistikou normovaného normálního rozdělení $z_{(i)}$ dle $x_{(i)} = \mu + \tau \left[\frac{\exp(\sigma z_{(i)}) - 1}{\sigma} \right] \approx \mu + \tau g_i(\sigma)$



V řadě případů je však možná standardizace

$s = (x - Q)/R$, kde R je parametr rozptýlení.

V případě shody obou rozdělení pak resultuje přímková závislost

$$x_{(i)} = Q + R \,.\, Q_S(P_i) = a + b \,.\, Q_S(P_i).$$

Odhad střední hodnoty odpovídá absolutnímu členu a
odhad směrodatné odchylky směrnicí b regresní přímky.

Pro odhad parametrů z Q-Q grafů je možno použít bud MNČ.

Blom doporučuje určení odhadu

$$g_i(\sigma) = \left[\frac{\exp(\sigma * \Phi^{-1}(P_i) - 1}{\sigma} \right],$$

kde je pořadová pravděpodobnost s korekcí $P_i = \frac{i + 0.3133\ \sigma - 0.3927}{N + 0.2146}$.

V případě známého σ je tedy možné odhadnout parametry μ, τ
jako úsek a směrnici regresní přímky $x_{(i)} \approx \mu + \tau g_i(\sigma)$.

Pro korekci lze určit prvky váhové matice $\boldsymbol{V}$ pro metodu vážených nejmenších čtverců ve tvaru

$$V_{ii} = 2(N+1)(N+2)h_i^2$$

$$V_{i+1,i} = V_{i,i+1} = -(N+1)(N+2)h_i h_{i+1}$$

$$V_{i,j} = 0 \text{ jinde,}$$

kde $h_i = \frac{\Phi\left[\Phi^{-1}(i/(N+1))\right]}{\exp[\sigma * \Phi^{-1}(i/(N+1))]}$.

Cenzorované výběry

Pro odhady parametrů v cenzorovaných výběrech lze požít jak metodu maximální věrohodnosti, tak i metody založené a pořádkových statistikách.

Cenzorování typu I: známe limitu detekce x_L (mez pod kterou se zaznamenává pouze „přítomnost" měření) a předpokládejme, známe rozdělení dat charakterizované hustotou pravděpodobnosti $f(x)$ resp. distribuční funkci $F(x)$.

Pro cenzorovaná měření lze při znalosti distribuční funkce měření určit pouze pravděpodobnost s jakou leží pod mezí detekce, která je rovna $F(x_L)$.

Veličina n_1 se uvažuje jako náhodná, protože nelze a priori předpovědět, v kterém vzorku dojde k překročení meze detekce a v kterém ne.

Všechny možné kombinace n_1 prvků, které ve výběru velikosti N leží pod limitou detekce jsou dány binomickým koeficientem $N!/(n_1!\,(N - n_1)!)$.

Věrohodnostní funkce má pro tento případ tvar

$$\ln(L) = \frac{N!}{n_1! * (N - n_1)!} F(X_L)^{n_1} * \prod_{i=n_1+1}^{N} f(x_{(i)}).$$

Pro případ normálního rozdělení dat nalezl Cohen vztahy:

pro odhad střední hodnoty $\bar{x}_C$ a rozptylu s_C^2
odpovídající maximalizaci věrohodnostní funkces využitím odhadů z necenzorované části dat

$$\bar{x}_N = \frac{1}{N - n_1} \sum_{i=n_1+1}^{N} x_{(i)} \quad \text{a} \quad s_N^2 = \frac{1}{N - n_1 - 1} \sum_{i=n_1+1}^{N} (x_{(i)} - \bar{x}_N)^2.$$

Platí, že $\bar{x}_C = \bar{x}_N - \lambda * (\bar{x}_N - x_L)$ a $s_C^2 = s_N^2 + \lambda * (\bar{x}_N - x_L)^2$

Parametr λ závisí na odhadnutém podílu cenzorovaných dat $h = n_1 / N$
a na parametru $g = s_N^2 / (\bar{x}_N - x_L)^2$.

Hodnoty λ jsou tabelovány a existují také empirické vztahy:

zanedbání cenzorované části dat vede k nadhodnocení aritmetického průměru a podhodnocení rozptylu v závislosti na velikosti parametru λ.

Postačuje **jednokrokový odhad** založený na předpokladu, že počet hodnot pod limitou detekce má binomické rozdělení a pro **odhad střední hodnoty** $\bar{x}_{CJ}$ a **rozptylu** s_{CJ}^2 pak platí

$$\bar{x}_{CJ} = \bar{x}_N - q * s_N \quad \text{a} \quad s_{CJ}^2 = \frac{\sum_{i=n_1+1}^{N} x_{(i)}^2}{N - n_1} - (\bar{x}_N)^2 - s_N^2 * (q * \Phi^{-1}(h) - q^2)$$

Korekční faktor q má tvar $q = \frac{N}{(N - n_1)\sqrt{2\pi}} \exp(-0.5 * [\Phi^{-1}(h)]^2)$.

Odhady $\bar{x}_{CJ}$ a s_{CJ}^2 lze tedy určit relativně snadno bez nutnosti použití speciálních tabulek.

Pro dvou-parametrové logaritmicko-normální rozdělení stačí místo hodnot $x_{(i)}$ použít jejich logaritmů $\ln(x_{(i)})$ a logaritmovat i limitu detekce.

Praktické doplňky

Při zpracování experimentálních dat záleží na množství informací, které jsou:

I. **Víme vše**: známe pravděpodobnostní model a stačí pouze ověření předpokladů před konfirmativní statistickou analýzou

II. **Nevíme nic**: postavíme datově závislý pravděpodobnostní model a provede se komplexní analýza dat (1. EDA průzkumová, 2. Ověření předpokladů, 3. Transformace, 4. Porovnání výběrového rozdělení s teoretickými).

III. **Něco víme**: postavíme empirický model se známými tak i datově závislými informacemi, pak se provede 1., 2., 3. a 4. analýza dat)

Odhad intervalu spolehlivosti střední hodnoty výběru $(x_1, x_2, \ldots x_N)$ z (ne)známého rozdělení $f(x)$.
Základní problém je nenulová šikmost ($g_1 \# 0$) a špičatost odpovídající nenormálnímu rozdělení ($g_2 \# 3$).

Doporučené další postupy:

1) **Robustní metody,**
2) **Využití zešikmených rozdělení,**
3) **Počítačově intenzivní metody,**
4) **Generalizovaná lineární regrese.**

Příklad 1. *Určení koncentrace nečistot v surovině*

Byla sledována koncentrace nečistot v μg/g: **DL, DL, 1.24, 1.49, 1.50, 1.56, 1.61, 1.78,**
kde **DL** značí hodnoty pod limitou detekce $x_D = 1$ μg/g.

Cíl: odhad střední hodnoty rozptylu a intervalu spolehlivosti za předpokladu normálního rozdělení.

Řešení:

1. metoda: Postup s vynecháním hodnot pod mezí detekce (nevhodné, chybné!)
Průměr = 1.53 a výběrová směrodatná odchylka $s = 0.18$.
Kvantil t rozdělení $t_{0.975}(5) = 2.571$ a 95% interval spolehlivosti ***UC* = 1.72, *LC* = 1.34.**

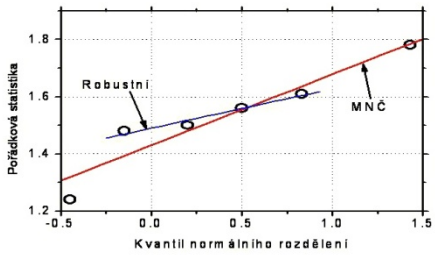
2. metoda: Maximalizace věrohodnostní funkce
Parametr h = 0.25, parametr g = 0.11 a tabelovaná hodnota $\lambda = 0.3387$. **Průměr = 1.35** a výběrová směrodatná odchylka $s = 0.36$ a 95% interval spolehlivosti ***UC* = 1.72, *LC* = 0.98.**

3. metoda: Jednokroková aproximace maximalizace věrohodnostní funkce
Průměr = 1.46 a výběrová směr. odchylka $s = 0.20$ a 95% interval spolehlivosti ***UC* = 1.67, *LC* = 1.25.**

4. metoda: Pořádkové statistiky
Ze směrnice a úseku určených klasickou MNČ vyšlo: **Průměr = 1.43** a výběrová směrodatná odchylka $s = 0.24$ a 95% interval spolehlivosti ***UC* = 1.68, *LC* = 1.18.**

Na obrázku je rankitový graf spolu s regresními přímkami pro **klasickou** a **robustní MNČ**.

Na obrázku je rankitový graf spolu s regresními přímkami pro **klasickou** a **robustní MNČ**.



Závěr: Je patrné, že **postupy berouci v úvahu limitu detekce** vedou k výrazně nižší dolní mezi intervalu spolehlivosti.

Formální aparát statistiky resp. přizpůsobení dat potřebám statistické analýzy **bez hlubšího rozboru** zde může vést ke katastrofálním závěrům.

Literatura

[1] Meloun M., Militký J.: *Zpracování experimentálních dat*, East Publishing Praha 1998
[2] Cohen, A.C.: Technometrics 3, 535 (1961)
[3] Boos D.D. , Hughes-Oliver J. M.: Amer. Statist. **54**, 121 (2000)
[4] Hall, P.: J.R. Stat. Sor. 54, 221 (1992)
[5] Chen L. : Environmetrica **6**, 181 (1995)
[6] Chen L.: J. Appl. Statist. **25**, 739 (1998)
[7] Smothers D.D. a kol.: Veterinary Parasditology **81**,211 (1999)
[8] May J. a kol. : Decision Sciences **31**, 129 (2000)
[9] Munro A. H., Wixley R. A.J. : J. Amer. Statist. Assoc. **65**, 212 (1970)
[10] Huybrechts T., a kol. : J. of Chromatography **A975**, 123 (2002)
[11] Haas C.N., Sheff P. A.: Environ. Sci. Technol. **24**, 912 (1990)
[12] Kuttatharmmakul S. a kol. : Trends in analytical chemistry 19, 215 (2000)
[13] Tiku M.L.: J. Amer. Statist. Assoc. **63**, 134 (1968)
[14] Shumway R. H. a kol.: Technometrics, **31**, 347-356 (1989)

Hornův postup malých výběrů

Analýza malých výběrů

- Závěry jsou vždy zatíženy značnou mírou nejistoty.
- Malých rozsahů jen tam, kde není možné zvýšit počet.

***n = 2*: 100(1 - α)%ní konfidenční interval střední hodnoty**

$$\frac{x_1 + x_2}{2} - T_\alpha \frac{|x_1 - x_2|}{2} \le \mu \le \frac{x_1 + x_2}{2} + T_\alpha \frac{|x_1 - x_2|}{2}$$

- pro normální rozdělení $T_\alpha = \mathrm{cotg}(\alpha\,\pi / 2)$, $T_{0.05} = 12.71$,
- pro rovnoměrné rozdělení $T_\alpha = 1/\alpha - 1$, tj. $T_{0.05} = 19$.

***n = 3*: 100(1 - α)%ní konfidenční interval střední hodnoty**

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

- pro normální rozdělení je $T'_\alpha \approx 1/\sqrt{\alpha} - 3\sqrt{\alpha}/4$ ..., $T'_\alpha = 4.30$.
- pro rovnoměrné rozdělení je $T_{0.05} = 5.74$,

***4 ≤ n ≤ 20*, (Hornův postup):**

je založený na pořádkových statistikách.

Hloubka pivotu je $H = (\mathrm{int}((n + 1)/2)+1)/2$

nebo $H = (\mathrm{int}((n + 1)/2 + 1)/2$,

Dolní pivot je $x_D = x_{(H)}$ a **horní pivot** $x_H = x_{(n+1-H)}$.

Odhadem parametru polohy je **pivotová polosuma**

$$P_L = \frac{x_D + x_H}{2}$$

odhadem parametru rozptýlení **pivotové rozpětí**

$$R_L = x_H - x_D$$

Náhodná veličina k testování

$$T_L = \frac{P_L}{R_L} = \frac{x_D + x_H}{2\,(x_H - x_D)}$$

má přibližně symetrické rozdělení, jehož vybrané kvantily jsou v tabulce.

95%ní interval spolehlivosti střední hodnoty se vypočte

$$P_L - R_L\, t_{L,0.975}(n) \le \mu \le P_L + R_L\, t_{L,0.975}(n)$$

Tabulka 3.11 Kvantily $t_{L,1-\alpha}(n)$ rozdělení T_L

n \ 1 - α	0.9	0.95	0.975	0.99	0.995
4	0.477	0.555	0.738	1.040	1.331
5	0.869	1.370	2.094	3.715	5.805
6	0.531	0.759	1.035	1.505	1.968
7	0.451	0.550	0.720	0.978	1.211
8	0.393	0.469	0.564	0.741	0.890
9	0.484	0.688	0.915	1.265	1.575
10	0.400	0.523	0.668	0.878	1.051
11	0.363	0.452	0.545	0.714	0.859
12	0.344	0.423	0.483	0.593	0.697
13	0.389	0.497	0.608	0.792	0.945
14	0.348	0.437	0.525	0.661	0.776
15	0.318	0.399	0.466	0.586	0.685
16	0.299	0.374	0.435	0.507	0.591
17	0.331	0.421	0.502	0.637	0.774
18	0.300	0.380	0.451	0.555	0.650
19	0.288	0.361	0.423	0.502	0.575
20	0.266	0.337	0.397	0.464	0.519

Úloha C3.11 *Test správnosti koncentrace tenzidů (Horn)*
Standardní vzorek obsahuje 2.5 mg/l anionaktivních tenzidů. Aplikujte i Hornův postup. Testujte, zda výsledky koncentrace standardu jsou správné. Jde o symetrické rozdělení ?
Data: Koncentrace tenzidů [mg/l]: 2.36 2.40 2.48 2.50 2.57 2.62 2.68
[Výsledky: Gauss. rozd., $\bar{x} = 2.52, \bar{x}_R = 2.51, \tilde{x}_{0.5} = 2.50, s = 0.12,$
$\hat{g}_1 = 0.04, \hat{g}_2 = 1.78, \quad 2.41 < \bar{x} < 2.62$*]*

Hornův postup:

1. Pořádkové statistiky:

i	1	2	3	4	5	6	7
$x_{(i)}$	2.36	2.40	2.48	2.50	2.57	2.62	2.68

2. Hloubka pivotu: n = 7, liché

$$H = integer\frac{\frac{n+1}{2}}{2} = \qquad \mathrm{int}(2.0) \approx 2$$

3. Pivoty:
Dolní pivot $x_D = x_{(H)} =$ $x_{(2)}$ = **2.40**
Horní pivot $x_H = x_{(n+1-H)} =$ $x_{(6)}$ = **2.62**

4. Pivotová polosuma $P_L = \frac{x_D + x_H}{2} =$ = **2.51**

5. Pivotové rozpětí $R_L = x_H - x_D =$ 2.62 - 2.40 = **0.22**

6. 95%ní interval spolehlivosti střední hodnoty μ: $t_{L,\,1-\alpha/2}(7) = 0.720$

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

$$2.51 - 0.22 \times 0.72 \le \mu \le 2.51 + 0.22 \times 0.72$$

$$\mathbf{2.35 \le \mu \le 2.67}$$