Nejlepší odhady polohy a rozptýlení chemických dat

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice
email: *milan.meloun@upce.cz, http://meloun.upce.cz*
a
Prof. Ing. Jiří Militký, CSc.,
Katedra textilních materiálů, Technická univerzita Liberec, 461 17 Liberec
email: *jiri.militky@vslib.cz*

Souhrn: *Pokud nemá být statistická analýza pouhým numerickým vyčíslením bez hlubšího smyslu, je pochopitelně třeba, aby byly ověřeny všechny předpoklady, které by vedly k návrhu daného postupu statistické analýzy. Při rutinním zpracování experimentálních dat se obyčejně provádí 1. popisná analýza s odhadem parametru polohy, rozptýlení a tvaru, 2. určení intervalů spolehlivosti a 3. testace významnosti parametrů. Při klasické popisné analýze se předpokládá splnění všech předpokladů o výběru, že totiž (a) určení vhodného rozsahu výběru, (b) jednotlivé prvky výběru* x_i *jsou vzájemně nezávislé, (c) výběr je homogenní a všechna* x_i *pocházejí ze stejného rozdělení, (d) rozdělení je gaussovské - normální.*

Jedním ze základních kroků interaktivní analýzy jednorozměrných dat je ověření základních předpokladů o výběru. Zatímco v prvním kroku analýzy dat se v exploratorní analýze užívá především grafických technik, jejichž diagnostiky jsou jednoduché a málo ovlivněné případnými vybočujícími měřeními čili heterogenitou dat v druhém kroku se v konfirmatorní analýze provádí statistické testování uvedených předpokladů. Je třeba mít na paměti, že malé porušení předpokladu normality nemusí být katastrofické s ohledem na výsledek analýzy. V zásadě platí, že pokud data nesplňují předpoklad normality, je v řadě případů možné zlepšit jejich rozdělení vhodnou transformací. V této práci budou ukázány přístupy rozličného často užívaného software k analýze jednorozměrných dat a poukázáno na nebezpečí získání falešných výsledků.

Teoretická část

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

A. V průzkumové analýze dat se vyšetřují *statistické zvláštnosti dat,* jako je

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

B. Pro případ rutinních měření se ověří *základní předpoklady,* kladené na výběr, jako jsou nezávislost prvků, homogenita výběru, dostatečný rozsah výběru a rozdělení výběru. Jsou-li závěry tohoto kroku optimistické, následuje vyčíslení klasických odhadů polohy a rozptýlení, tj. obyčejně aritmetického průměru a rozptylu. Dále se vyčíslí intervaly spolehlivosti, následované testováním statistických hypotéz. V pesimistickém případě následuje další pokus o úpravu dat.

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

Postup statistické analýzy jednorozměrných dat, prováděné v interaktivním režimu na počítači, lze shrnout do bloků, i když lze jednotlivé bloky provádět samostatně: blok A, blok B, blok A+B, blok B+C a konečně všechny bloky A+B+C.

Průzkumová (exploratorní) analýza dat EDA

Prvním krokem v analýze jednorozměrných dat je průzkumová čili exploratorní analýza. Vychází se z *pořádkových statistik* výběru, tj. z prvků výběru uspořádaných vzestupně $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Platí, že střední hodnota i-té pořádkové statistiky $E(x_{(i)})$ je rovna 100 P_i procentnímu kvantilu výběrového rozdělení $Q(P_i)$ a symbol $P_i \approx i/(n + 1)$ označuje *pořadovou pravděpodobnost*. V průzkumové analýze se často používá speciálních *kvantilů L* pro pořadové pravděpodobnosti $P_i = 2^{-i}$, $i = 1, 2, \ldots$, které se také nazývají *písmenové hodnoty*.

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

Symbol u_{P_i} označuje kvantil normovaného normálního rozdělení $N(0, 1)$. Kromě *mediánu* ($i = 1$) existují pro každé $i > 1$ dvojice kvantilů, a to *dolní* a *horní písmenová hodnota* L_D a L_H. Dolní písmenová hodnota je pro pořadovou pravděpodobnost $P_i = 2^{-i}$, zatímco horní je pro $P_i = 1 - 2^{-i}$. Mezi základní statistické zvláštnosti rozdělení dat patří symetrie výběrového rozdělení a jeho relativní délky konců ve srovnání s normálním rozdělením. K vyjádření symetrie a špičatosti v různých vzdálenostech od mediánu se užívají jednoduché diagnostické grafy a diagramy:

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

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.

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.

Houslový diagram (osa x: název výběru proměnné, osa y: percentily, hodnoty proměnné). Diagram je kombinací krabicového grafu a dvou vertikálních, zrcadlově k sobě zobrazených grafů hustoty. Jeden graf hustoty roste směrem doprava a druhý doleva. Diagram zobrazuje píky a údolí stejně jako graf hustoty pravděpodobnosti. Medián je zobrazen černým kolečkem a začátek a konec úsečky zobrazuje dolní a horní kvantil.

Kvantilový graf (osa x: pořadová pravděpodobnost P_i, osa y: pořádková statistika $x_{(i)}$). Umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, které může být symetrické, zešikmené k vyšším nebo nižším hodnotám.

Krabicový graf (osa x: úměrná hodnotám x, osa y: interval úměrný hodnotě $\sqrt{n}$). Pro částečnou sumarizaci dat lze využít krabicového grafu, který umožňuje znázornění robustního odhadu polohy, mediánu M, dále posouzení symetrie v okolí kvartilů, posouzení symetrie u konců rozdělení a konečně identifikaci odlehlých dat. Krabicový graf je obdélník o délce $R_F = F_H - F_D$ s vhodně zvolenou šířkou, která je úměrná hodnotě $\sqrt{n}$. V místě mediánu M je vertikální čára. Od obou protilehlých stran tohoto obdélníku pokračují úsečky. Prvky výběru, ležící mimo

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

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

Graf polosum (osa x: pořádkové statistiky $x_{(i)}$, osa y: $Z_i = 0.5\ (x_{(n+1-i)} + x_{(i)})$). Pro symetrické rozdělení je grafem polosum horizontální přímka, určená rovnicí $y = M$. U tohoto grafu je důležité, že zde body oscilují okolo horizontální přímky a vykazují tak náhodný shluk (mrak) a měřítko osy y je silně detailní. Naopak, asymetrické rozdělení vykazuje nenáhodný trend a body pak neoscilují okolo horizontální přímky a měřítko osy y není detailní.

Graf rozptýlení s kvantily (osa x: P_i, osa y: $x_{(i)}$). Základem je odhad kvantilové funkce výběru, který se získá spojením bodů $\{x_{(i)}, P_i\}$ lineárními úseky. Pro symetrická rozdělení má kvantilová funkce sigmoidální tvar. Pro rozdělení zešikmená k vyšším hodnotám je konvexně rostoucí a pro rozdělení zešikmená k nižším hodnotám konkávně rostoucí. Do grafu se zakreslují tři pomocné kvantilové obdélníky: (a) kvartilový obdélník F. (b) oktilový obdélník E. (c) sedecilový obdélník D.

Histogram (osa x: proměnná x, osa y: úměrná hustotě pravděpodobnosti). Jde o obrys sloupcového grafu, kde jsou na ose x jednotlivé třídy, definující šířky sloupců. Výšky sloupců odpovídají empirickým hustotám pravděpodobnosti. Kvalitu histogramu ovlivňuje ve značné míře volba počtu tříd L. Pro přibližně symetrická rozdělení výběru lze vyčíslit počet tříd L podle vztahu $L = \mathrm{int}(2\sqrt{n})$, kde funkce int($x$) označuje celočíselnou část čísla x.

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

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 $Q_T(P)$. Jako odhad kvantilové funkce výběru se využívají pořádkové statistiky $x_{(i)}$. Při shodě výběrového rozdělení se zvoleným teoretickým rozdělením platí přibližná rovnost kvantilů $x_{(i)} \approx Q_T(P_i)$, kde P_i je pořadová pravděpodobnost, a závislost $x_{(i)}$ na $Q_T(P_i)$ je přibližně přímka. Korelační koeficient r_{xy} je pak kritériem těsnosti proložení této přímky při hledání typu neznámého rozdělení.

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

Průběh průzkumové analýzy dat

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

A. Postup analýzy rutinních dat

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

a) testování nezávislosti prvků výběru - autokorelace,
b) testování homogenity výběru,
c) testování normality rozdělení výběru.

Z grafických metod se k předběžné analýze rutinních dat nejčastěji užívá *rankitového Q-Q grafu* a *grafu rozptýlení s kvantily*. Nejsou-li však o rozdělení dat dostupné žádné informace, nebo očekává-li se výrazně nenormální rozdělení, je vhodné provést

a) průzkumovou analýzu dat využitím řady grafických diagnostik,
b) určení výběrového rozdělení a jeho konstrukce.

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

B. Postup při nesplnění předpokladů o datech

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

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

3. Přítomnost vybočujících hodnot. Na základě logické analýzy je třeba nejdříve zvážit, zda nejde o zešikmené rozdělení. Body, které se jeví vybočující pro symetrické (speciálně normální) rozdělení, mohou být pro zešikmená 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 odlehlá 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 nebo robustní transformace. 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).

C. Transformace dat

Pokud se na základě analýzy dat zjistí, že rozdělení výběru dat se systematicky odlišuje od rozdělení normálního, vzniká problém, jak data vůbec vyhodnotit. Často je pak vhodná robustní transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě rozdělení a oslabení vlivu odlehlých hodnot.

1. *Stabilizace rozptylu* vyžaduje nalezení transformace $y = g(x)$, ve které je již rozptyl $\sigma^2(y)$ konstantní.

2. *Zesymetričtění rozdělení* výběru je možné provést užitím prosté *mocninné transformace*

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

Mocninná transformace však nezachovává měřítko, není vzhledem k hodnotě λ všude spojitá a hodí se pouze pro kladná data. Optimální odhad exponentu λ se hledá s ohledem na optimalizaci charakteristik asymetrie (šikmosti) a špičatosti. K určení optimálního λ lze užít i *rankitového grafu*, kdy pro optimální exponent λ budou kvantily $y_{(i)}$ ležet na přímce, nebo *selekčního grafu dle Hinese a Hinesové*. Podle umístění experimentálních bodů v okolí nomogramu teoretických křivek selekčního grafu lze vizuálně odhadovat velikost λ a posuzovat tak kvalitu transformace v různých vzdálenostech od mediánu.

Pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti je vhodná *Boxova-Coxova transformace*

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

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

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

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

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

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

Ilustrativní příklady

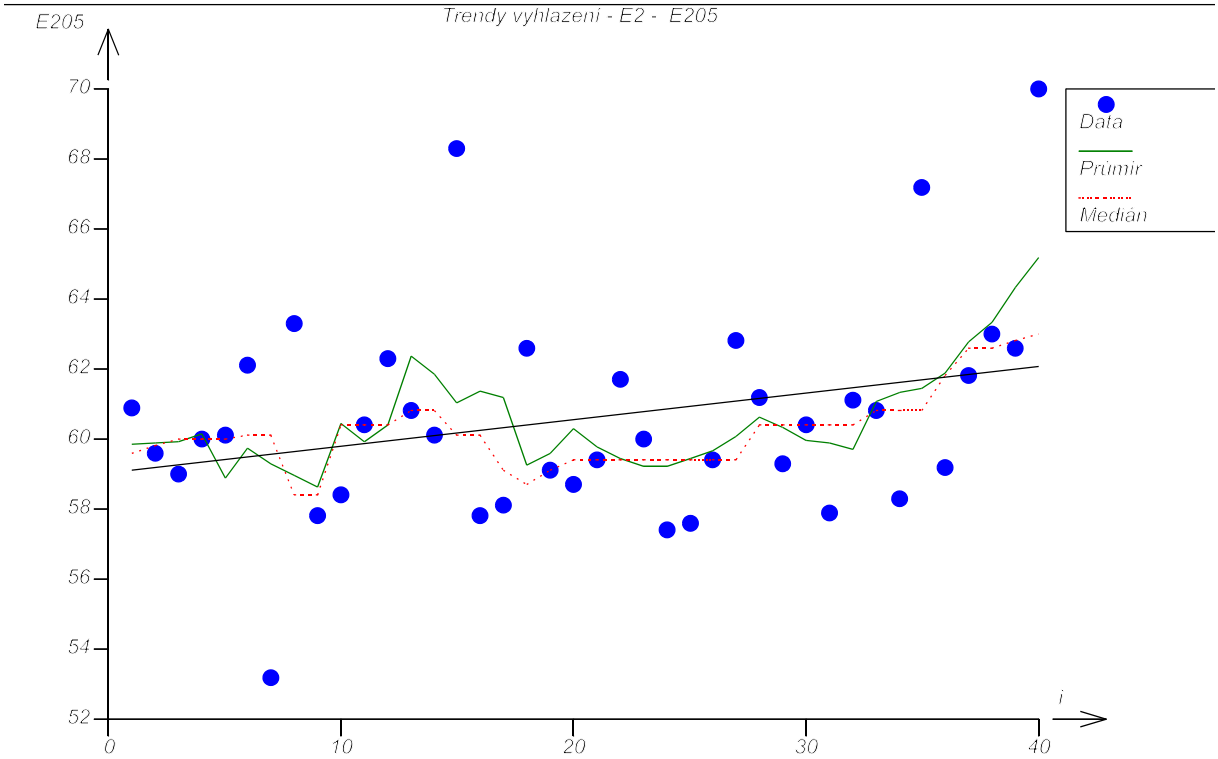
Úloha 1. *Podezřelé a odlehlé hodnoty obsahu PCB v oleji* (E205)
V referenčním vzorku oleje byl stanoven obsah PCB [mg/kg] (Delor 103, Delor 106). Jsou v datech podezřelé a odlehlé hodnoty? Určete střední hodnotu.

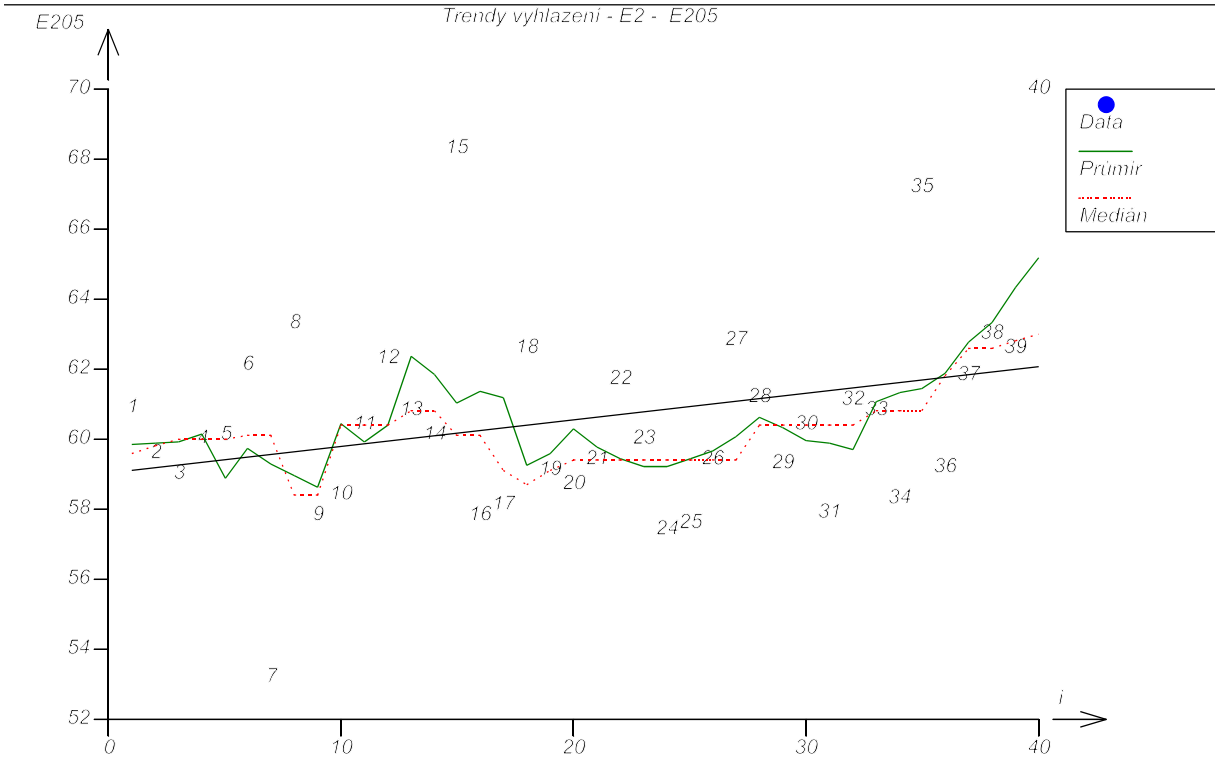
Úloha 2. *Statistické zvláštnosti výběru hodnot obsahu kadmia v bramborách* (E219). Na data o koncentraci kadmia v bramborách v oblastech jižní Moravy a Čech aplikujte průzkumovou analýzu dat. Jsou ve výběru odlehlé hodnoty? Vyčíslete střední hodnotu.

Úloha 3. *Stanovení kadmia v grahamových rohlících* (E224)
V mezilaboratorním testu stanovení kadmia v grahamových rohlících je třeba stanovit střední hodnotu. Nalezly některé laboratoře podezřelé a odlehlé hodnoty?

1. úloha: *Podezřelé a odlehlé hodnoty obsahu PCB v oleji,*
(software QC-Expert, ADSTAT2.5)

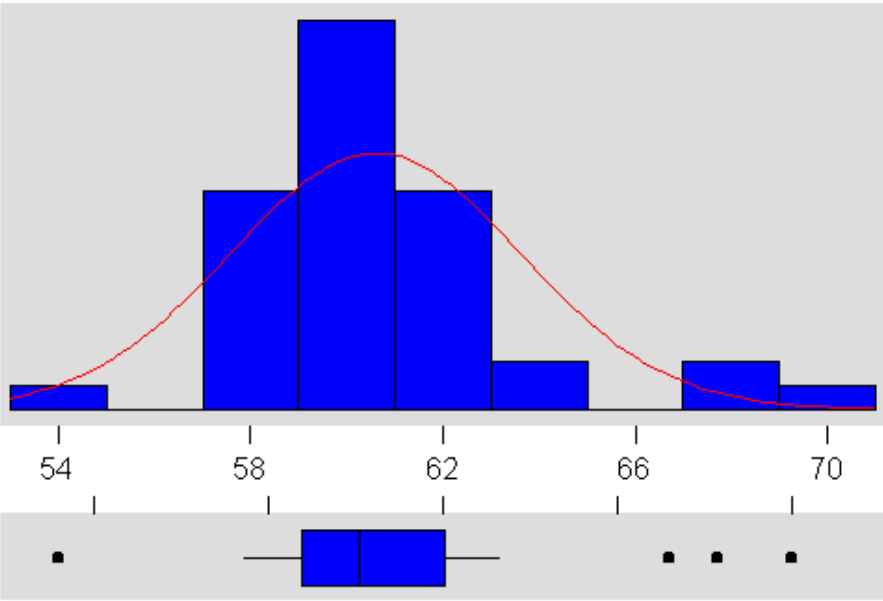
Indexový diagram trendu dat E205

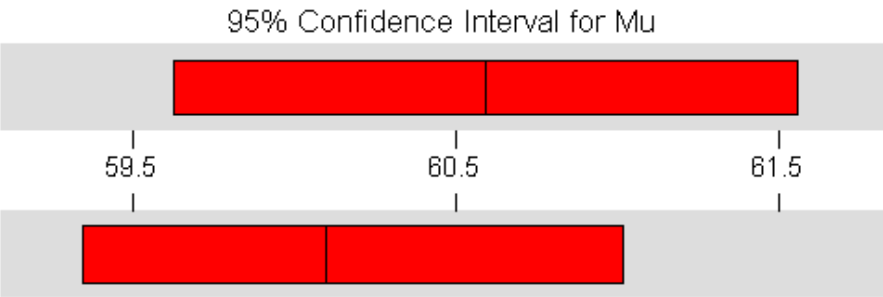




Minitab

Descriptive Statistics





Variable: E205

Anderson-Darling Normality Test

A-Squared:	1.267
P-Value:	0.002
Mean	60.5925
StDev	3.0224
Variance	9.13507
Skewness	1.03922
Kurtosis	2.81121
N	40
Minimum	53.2000
1st Quartile	58.7750
Median	60.1000
3rd Quartile	62.0250
Maximum	70.0000

95% Confidence Interval for Mu

59.6259	61.5591

95% Confidence Interval for Sigma

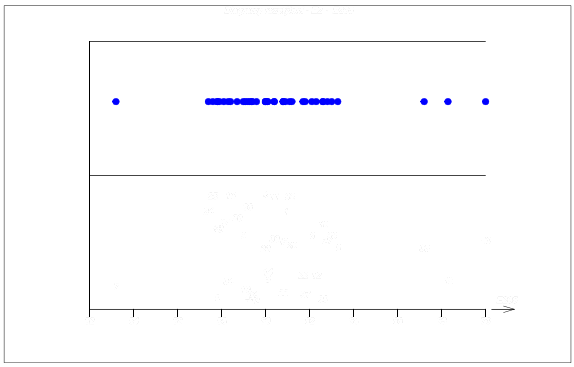
2.4759	3.8809

95% Confidence Interval for Median

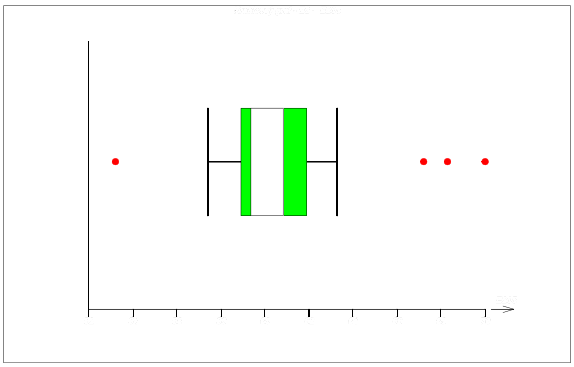
59.3411	61.0178

Exploratorní analýza dat E205

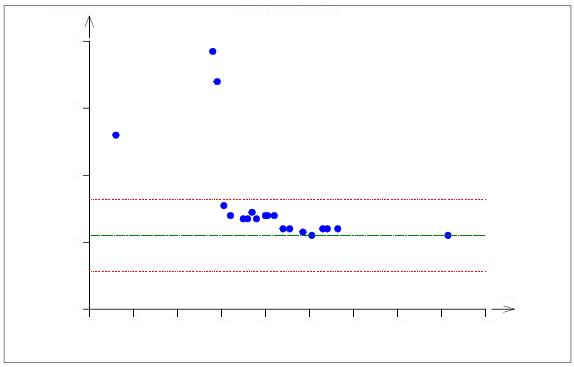
(software QC-Expert, ADSTAT2.5)



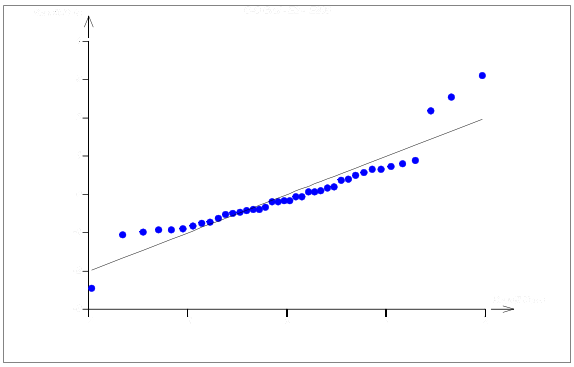
Rozmítnutý diagram rozptýlení



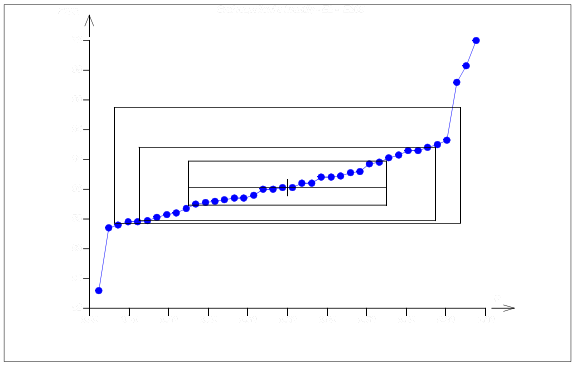
Krabicový graf



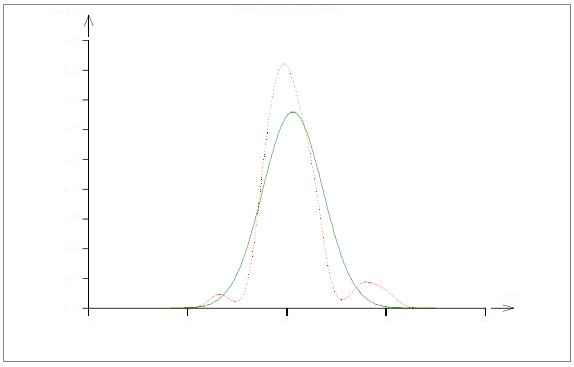
Graf polosum



Rankitový Q-Q graf



Graf rozptýlení s kvantily



Graf hustoty pravděpodobnosti

Diagnoza prvků výběru dat:

1. *Diagram rozptýlení a rozmítnutý diagram rozptýlení*: ukazuje na 3 odlehlé body v horní části diagramu a 1 odlehlý bod v dolní části diagramu.
2. *Vrubový krabicový graf*: jsou detekovány 4 odlehlé body. Téměř symetrická krabice je rozdělena na dvě části mediánem.
3. *Graf polosum*: indikuje část bodů jako vybočujících ze symetrického rozdělení.
4. *Rankitový Q-Q graf*: jelikož část bodů neleží na přímce, jde o asymetrické rozdělení.
5. *Graf rozptýlení s kvantily*: mírná asymetrie kvantilových obdélníků obdélníků dokazuje asymetrické rozdělení. Body ležící vně sedecilového obdélníku indikuje tato pomůcka jako odlehlé.
6. *Jádrový odhad hustoty pravděpodobnosti*: ve srovnání s normálním rozdělením je mírné sešikmení.

Odhady polohy, rozptýlení a tvaru E205

(software QC-Expert, ADSTAT2.5)

Průměr: 60.59 Spodní mez: 59.63 Horní mez: 61.56 Směr. odchylka: 3.02
Šikmost: 1.00 Odchylka od 0 je významná
Špičatost: 5.33 Odchylka od 3 je významná
Medián: 60.10 IS spodní:57.88 IS horní:62.32 Median. směr. odch.: 1.10

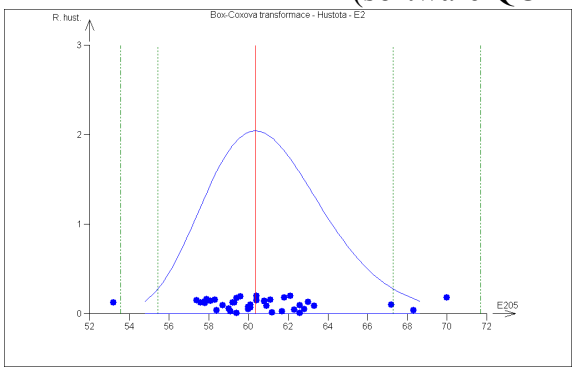
Testy výběrových předpokladů E205

(software QC-Expert, ADSTAT2.5)

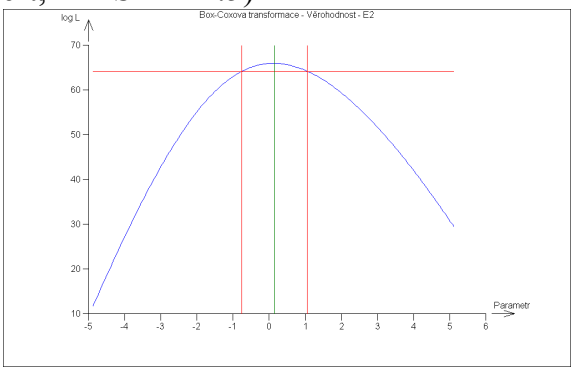
Znaménkový test: Data jsou nezávislá, Test normality: Normalita je přijata,Test homogenity: Homogenita je zamítnuta, Vybočujících hodnot: 1
Vnitřní hradby: Spodní: 52.00 Horní: 68.50

Transformace dat E205

(software QC-Expert, ADSTAT2.5)



Graf hustoty pravděpodobnosti



Graf maximální věrohodnost

Exponenciální transformace E205

(software QC-Expert, ADSTAT2.5)

Optimální parametr: 0.202
Opravený průměr: 60.33 Spodní mez: 59.44 Horní mez: 61.27

Box-Coxova transformace E205:

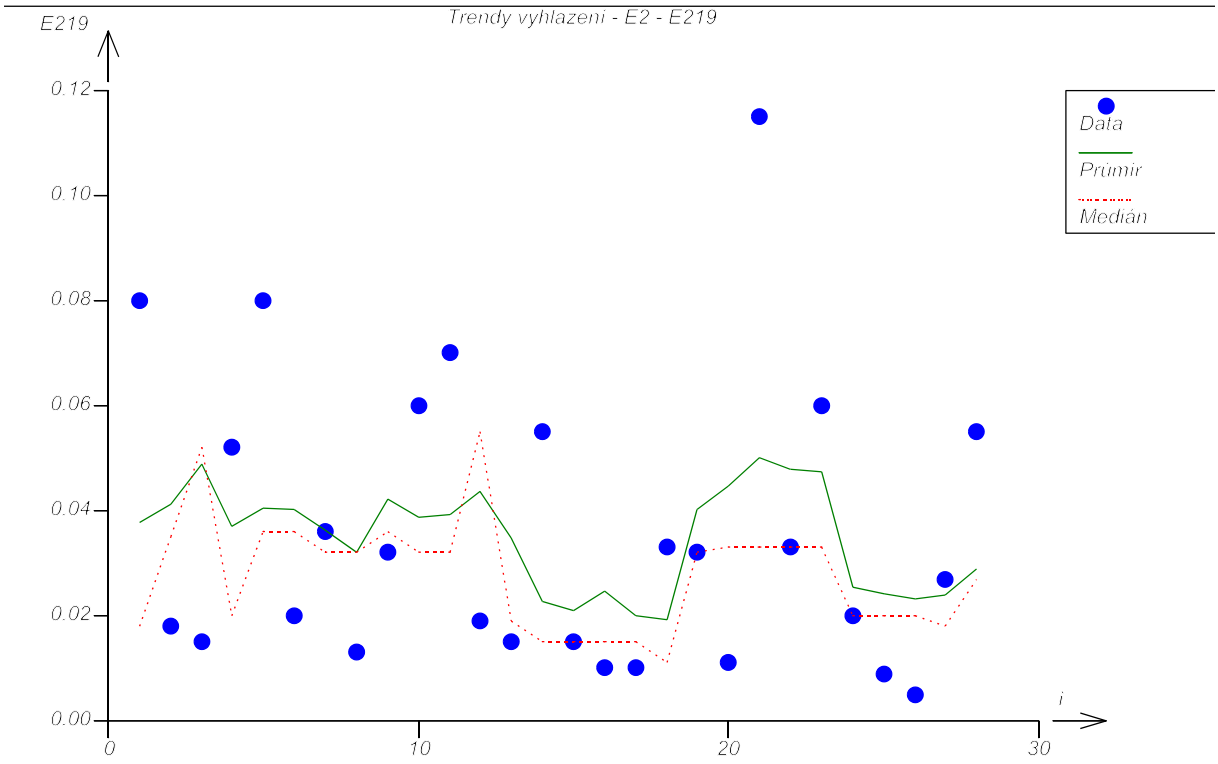
Optimální parametr: 0.124 Meze-dolní: -0.755 Horní mez: 1.054
Závěr: Transformace je statisticky nevýznamná.
Opravený průměr: 60.35

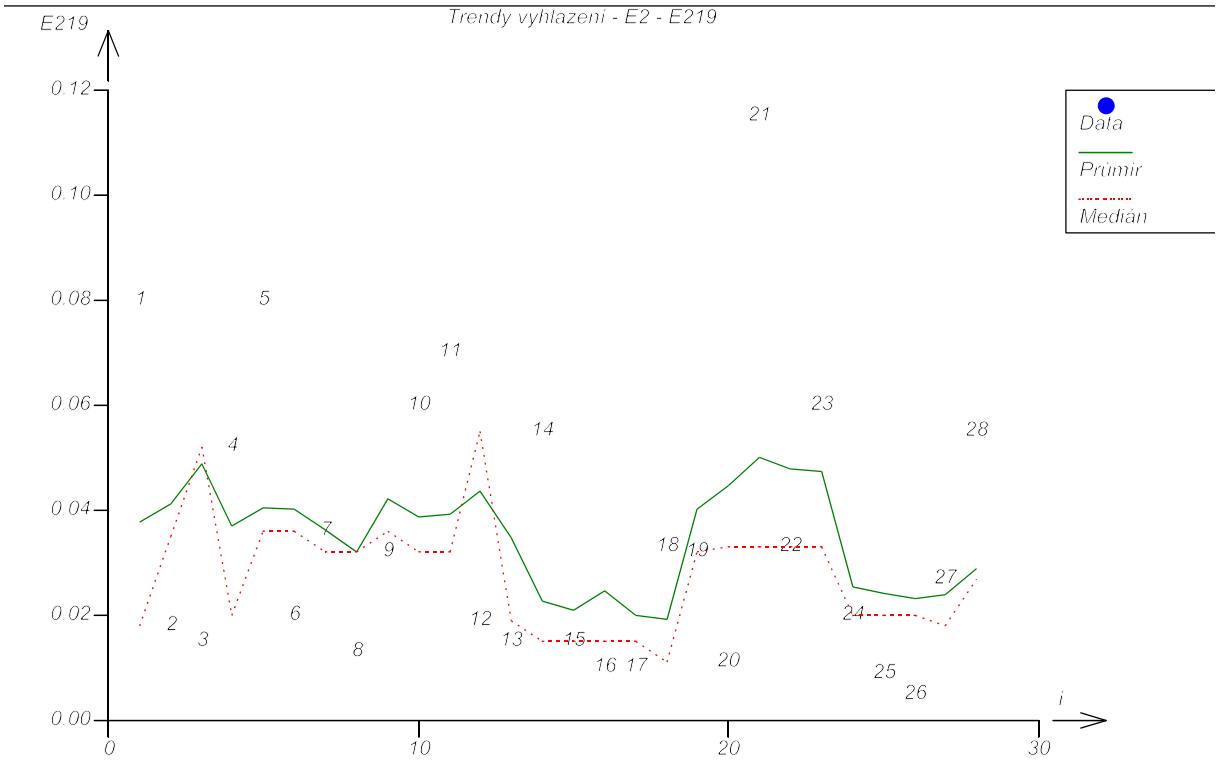
Závěr úlohy: za nejlepší odhad střední hodnoty je třeba považovat 60.35.

2. úloha: *Statistické zvláštnosti výběru hodnot obsahu kadmia v bramborách*

Indexový diagram trendu dat E219

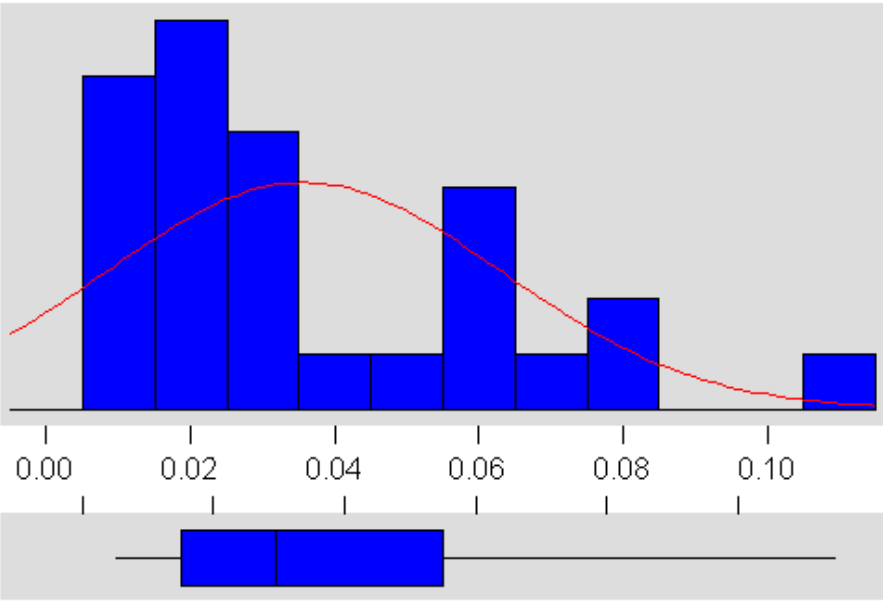
software QC-Expert (ADSTAT2.5)

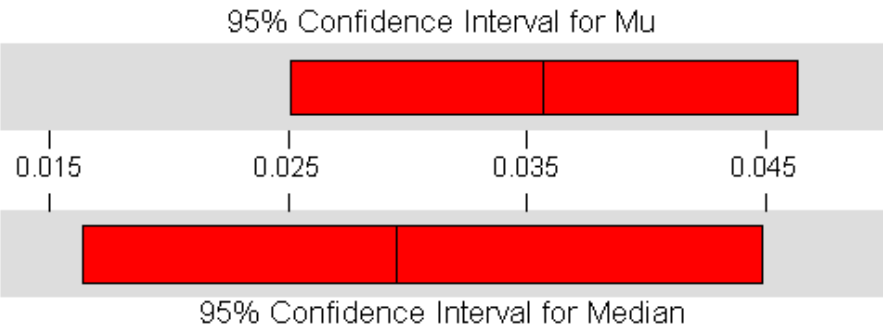




Minitab

Descriptive Statistics





Variable: E219

Anderson-Darling Normality Test

A-Squared:	1.201
P-Value:	0.003
Mean	3.57E-02
StDev	2.73E-02
Variance	7.47E-04
Skewness	1.18251
Kurtosis	1.06223
N	28
Minimum	0.005000
1st Quartile	0.015000
Median	0.029500
3rd Quartile	0.055000
Maximum	0.115000

95% Confidence Interval for Mu

0.025116	0.046313

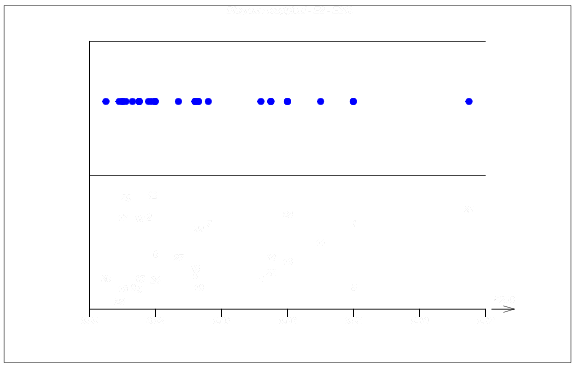
95% Confidence Interval for Sigma

0.021610	0.037204

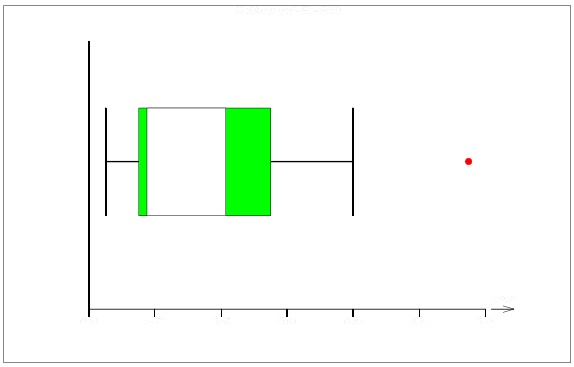
95% Confidence Interval for Median

0.016345	0.044827

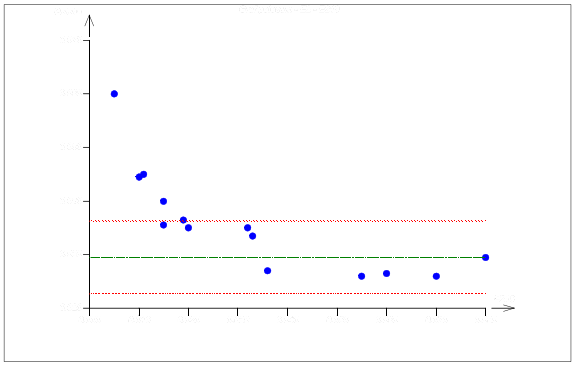
Exploratorní analýza E219



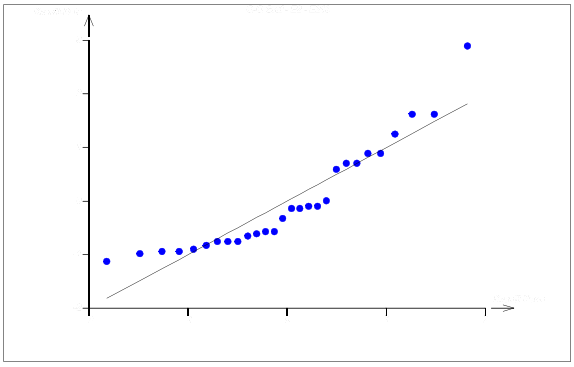
Rozmítnutý diagram rozptýlení



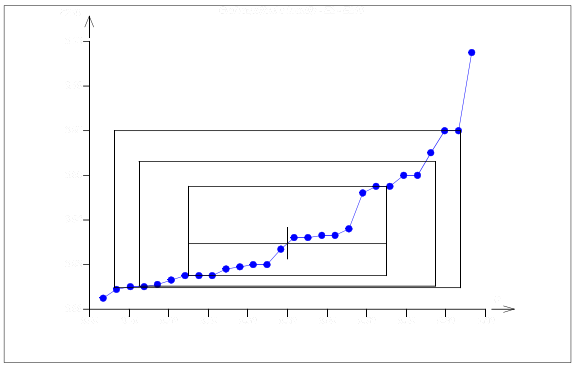
Krabicový graf



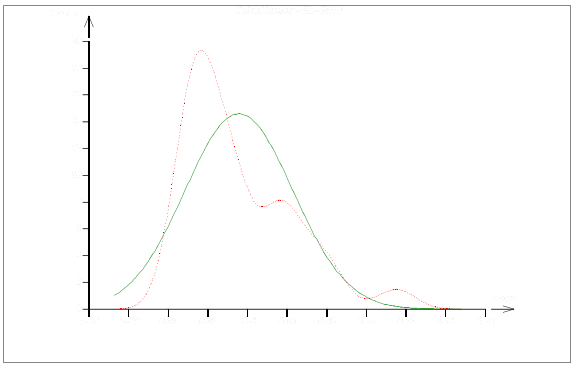
Graf polosum



Rankitový Q-Q graf



Graf rozptýlení s kvantily



Graf hustoty pravděpodobnosti

Diagnoza prvků výběru dat:

1. *Diagram rozptýlení a rozmítnutý diagram rozptýlení*: ukazuje na 1 až 2 odlehlé body v horní části.
2. *Vrubový krabicový graf*: je detekován 1 odlehlý bod. Asymetrická krabice je rozdělena na dvě části mediánem, rozdílná délka vousů ukazuje na asymetrii v koncích rozdělení.
3. *Graf polosum*: indikuje část bodů jako vybočujících ze symetrického rozdělení.
4. *Rankitový Q-Q graf*: jelikož část bodů neleží na přímce, jde o asymetrické rozdělení.
5. *Graf rozptýlení s kvantily*: asymetrie kvantilových obdélníků obdélníků dokazuje asymetrické rozdělení. Body ležící vně sedecilového obdélníku indikuje tato pomůcka jako odlehlé.
6. *Jádrový odhad hustoty pravděpodobnosti*: ve srovnání s normálním rozdělením je patrné sešikmení.

Odhady polohy, rozptýlení a tvaru E219

(software QC-Expert, ADSTAT2.5)

Průměr: 0.036 Spodní mez: 0.025 Horní mez: 0.046 Směr. odchylka: 0.027
Šikmost: 1.12 Odchylka od 0 je významná
Špičatost: 3.67 Odchylka od 3 je nevýznamná
Medián: 0.030 IS spodní: -0.007 IS horní: 0.067 Median. sm. odch.: 0.018

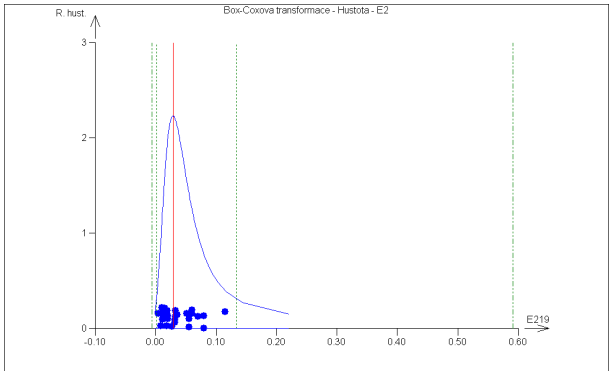
Testy výběrových předpokladů E219

(software QC-Expert, ADSTAT2.5)

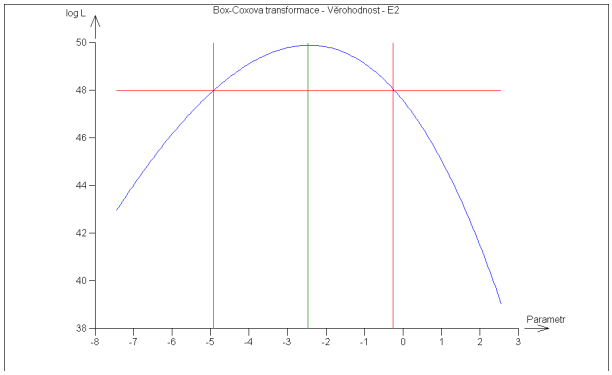
Znaménkový test: Data jsou nezávislá, Test normality: Normalita je přijata,
Test homogenity: Homogenita je přijata, Vybočujících hodnot: 0
Vnitřní hradby: Spodní: -0.070 Horní: 0.140

Transformace dat E219

(software QC-Expert, ADSTAT2.5)



Graf hustoty pravděpodobnosti



Graf maximální věrohodnosti

Exponenciální transformace E219

(software QC-Expert, ADSTAT2.5)

Optimální parametr: 1.074
Opravený průměr: 0.027 Spodní mez: 0.020 Horní mez: 0.035

Box-Coxova transformace E219

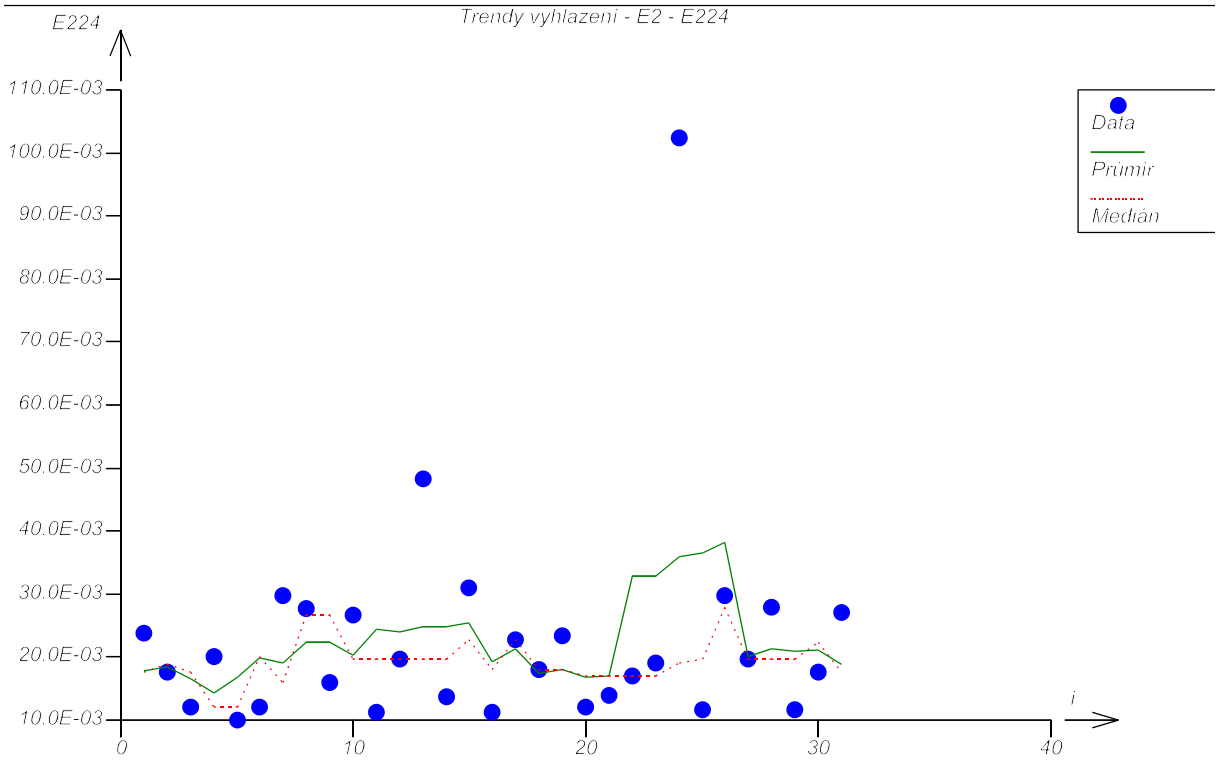
(software QC-Expert, ADSTAT2.5)

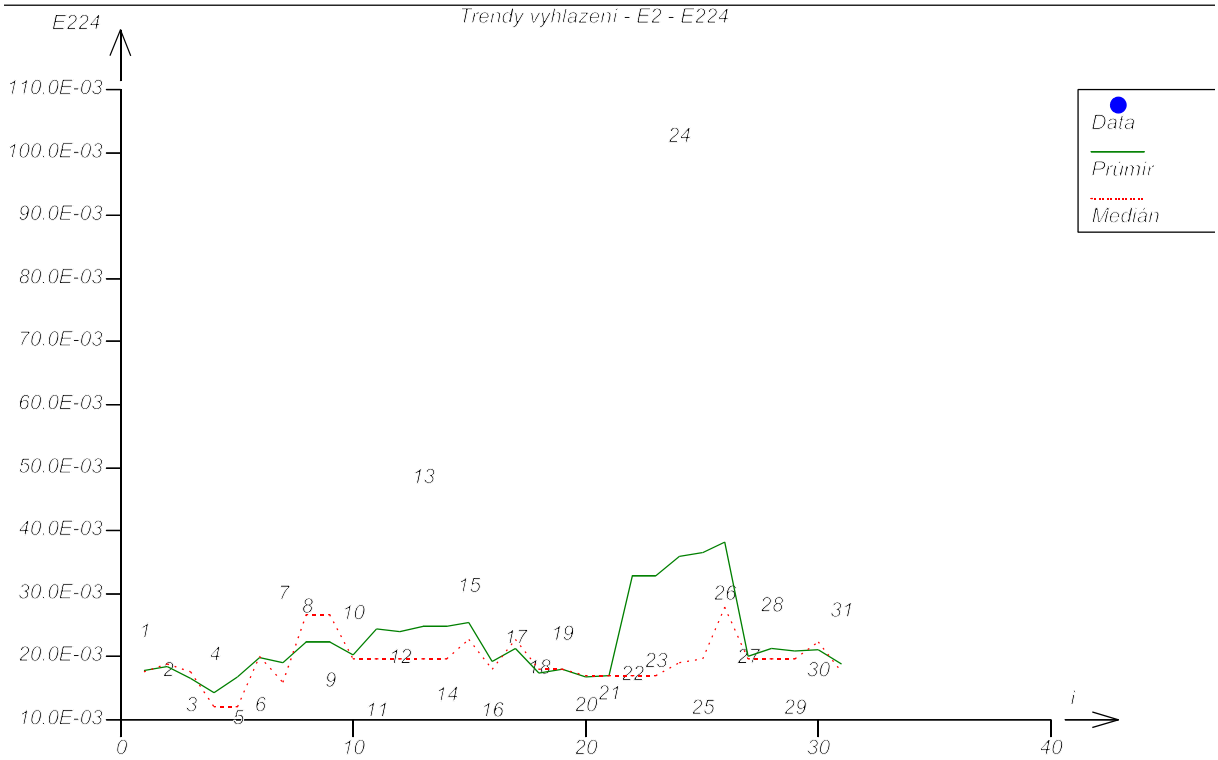
Optimální parametr: -2.440 Meze-dolní: -4.928 Horní mez: -0.254
Závěr: Transformace je statisticky významná.
Opravený průměr: 0.029

Závěr úlohy: za nejlepší odhad střední hodnoty je třeba považovat 0.029.

3. úloha: *Stanovení kadmia v grahamových rohlících E224*

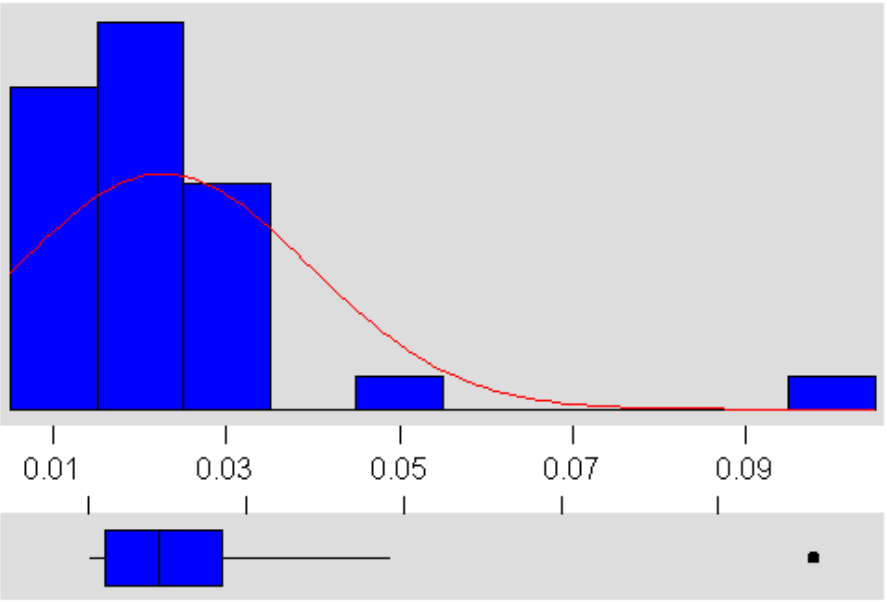
Indexový diagram trendu dat E224
software QC-Expert (ADSTAT2.5)

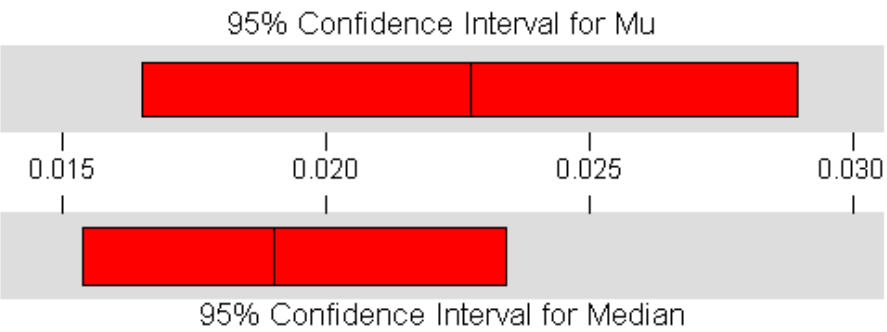




Minitab

Descriptive Statistics





Variable: E224

Anderson-Darling Normality Test

A-Squared:	3.323
P-Value:	0.000
Mean	2.27E-02
StDev	1.69E-02
Variance	2.87E-04
Skewness	3.73263
Kurtosis	16.8349
N	31
Minimum	0.010000
1st Quartile	0.012000
Median	0.019000
3rd Quartile	0.027000
Maximum	0.102300

95% Confidence Interval for Mu

0.016519	0.028939

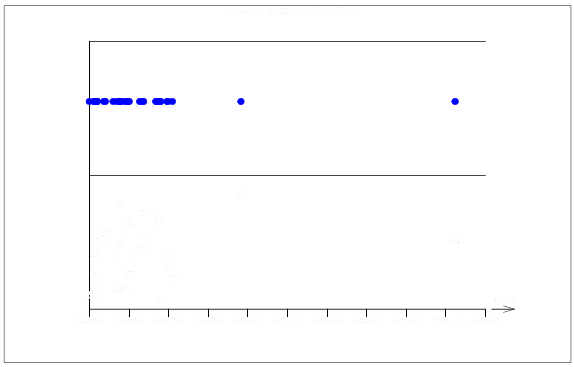
95% Confidence Interval for Sigma

0.013529	0.022630

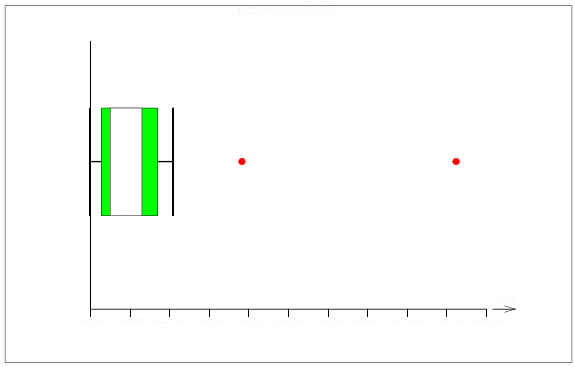
95% Confidence Interval for Median

0.015350	0.023430

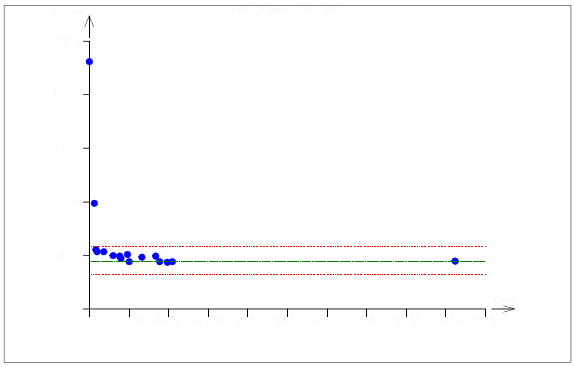
Exploratorní analýza E224



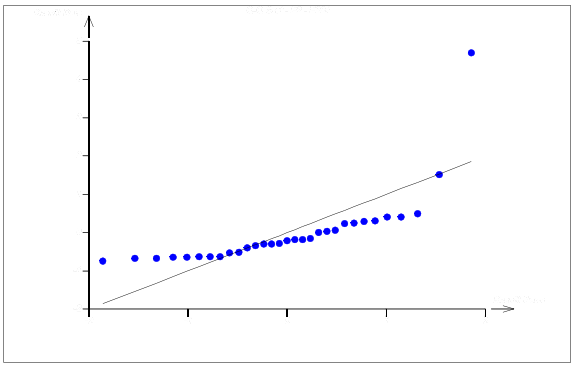
Rozmítnutý diagram rozptýlení



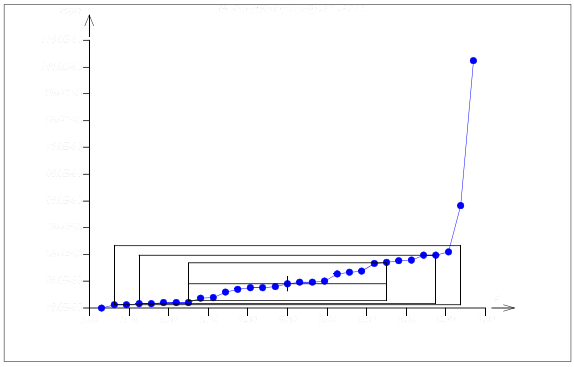
Krabicový graf



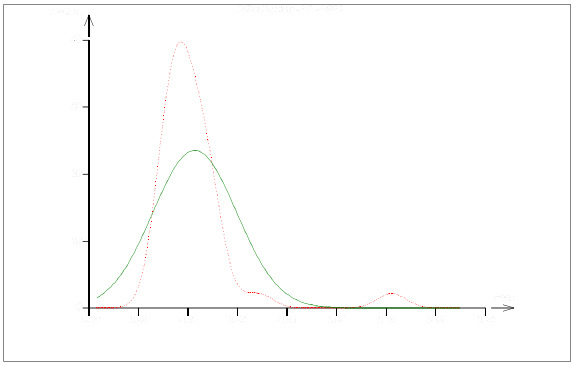
Graf polosum



Rankitový Q-Q graf



Graf rozptýlení s kvantily



Graf hustoty pravděpodobnosti

Diagnoza prvků výběru dat:

1. *Diagram rozptýlení a rozmítnutý diagram rozptýlení*: ukazuje na 2 odlehlé body v horní části.
2. *Vrubový krabicový graf*: jsou detekovány 2 odlehlé body. Téměř symetrická krabice je rozdělena na dvě části mediánem.
3. *Graf polosum*: indikuje poměrně málo bodů jako vybočujících ze symetrického rozdělení.
4. *Rankitový Q-Q graf*: jelikož část bodů neleží na přímce, jde o asymetrické rozdělení.
5. *Graf rozptýlení s kvantily*: mírná asymetrie kvantilových obdélníků obdélníků dokazuje asymetrické rozdělení. Dva body ležící vně sedecilového obdélníku indikuje tato pomůcka jako odlehlé.
6. *Jádrový odhad hustoty pravděpodobnosti*: ve srovnání s normálním rozdělením je patrná mírná asymetrie.

Odhady polohy, rozptýlení a tvaru E224

(software QC-Expert, ADSTAT2.5)

Průměr: 0.022 Spodní mez: 0.017 Horní mez: 0.029 Směr. odchylka: 0.017
Šikmost: 3.55 Odchylka od 0 je významná
Špičatost: 17.05 Odchylka od 3 je významná
Medián: 0.019 IS spodní: 0.010 IS horní: 0.028 Median. sm. odch.: 0.005

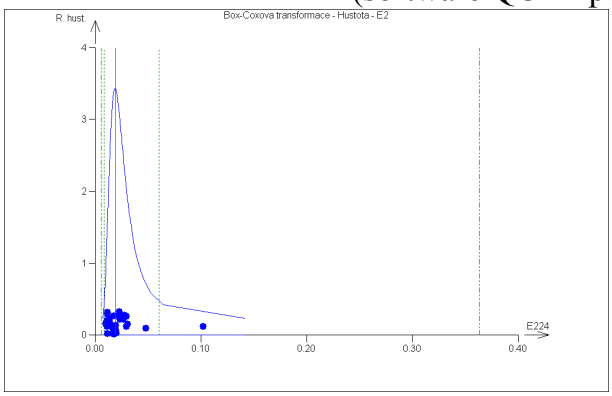
Testy výběrových předpokladů E224

(software QC-Expert, ADSTAT2.5)

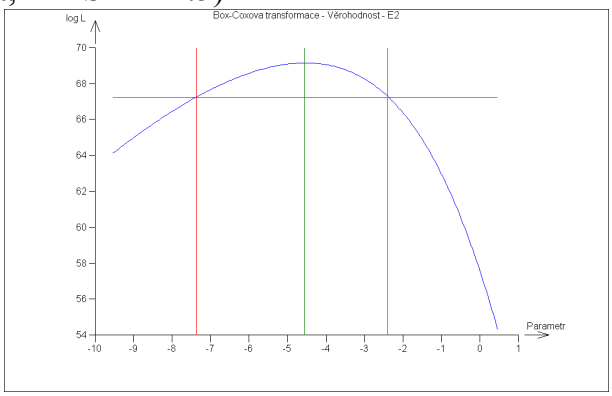
Znaménkový test: Data jsou nezávislá, Test normality: Normalita je zamítnuta,
Test homogenity: Homogenita je zamítnuta, Vybočujících hodnot: 1
Vnitřní hradby: Spodní: -0.019 Horní: 0.058

Transformace E224

(software QC-Expert, ADSTAT2.5)



Graf hustoty pravděpodobnosti



Graf maximální věrohodnosti

Exponenciální transformace E224

(software QC-Expert, ADSTAT2.5)

Optimální parametr: 1.858
Opravený průměr: 0.018 Spodní mez: 0.016 Horní mez: 0.021

Box-Coxova transformace E224

(software QC-Expert, ADSTAT2.5)

Optimální parametr: -4.534 Meze-dolní: -7.373 Horní mez: -2.398
Závěr: Transformace je statisticky významná.
Opravený průměr: 0.019

Závěr úlohy: za nejlepší odhad střední hodnoty je třeba považovat 0.019.

Závěr

Na základě předložených vyřešených úloh lze dospět k následujícímu obecnému postupu analýzy jednorozměrných dat:

Přehled operací analýzy jednorozměrných dat

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

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

B. Ověření předpokladů o datech:

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

C. 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.

Doporučená literatura:

1. M. Meloun, J. Militký: *Kompendium statistického zpracování dat*, Academia, 2002.
2. M. Meloun, J. Militký: *Statistické zpracování experimentálních dat,* East Publishing, 1996.