Postup analýzy dat

1. Průzkumová analýza dat:

Diagnostické grafy: *stupeň symetrie rozdělení*
lokální koncentrace dat
vybočující data

2. Ověření předpokladů výběru dat:

Diagnosticky, testy: *ověření normality*
ověření nezávislosti
ověření homogenity
určení minimální četnosti

3. Transformace dat:

Analýza dat: *originální data*
data po mocninné transformaci
data po Box-Coxově transformaci

4. Parametry polohy, rozptýlení a tvaru:

Analýza 1 výběru:
klasické odhady - *průměr*
- *rozptyl*

robustní odhady - *medián*
- *uřezané průměry*
- *winsorizovaný rozptyl*
- *interkvantilové rozpětí*

adaptivní odhady

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

Identifikace statistických zvláštností výběru dat

(1) Stupeň symetrie rozdělení výběru

(2) Stupeň špičatosti rozdělení výběru

(3) Lokální koncentrace dat

(4) Přítomnost vybočujících hodnot (měření)

2. kapitola

EDA: Exploratorní analýza dat

Řešení úloh z Kompendia:

B204

Úloha B2.04 *Typ rozdělení a odlehlé body obsahu ergosterinu v kalciferolu*

Zadání: Při výrobě kalciferolu se provádí kontrola meziproduktu 3,5 DNB esteru kalciferolu metodou HPLC. Sleduje se obsah přítomného ergosterinu jako nečistoty, jehož střední hodnota by neměla přesáhnout 0.4 %.

(1) Metodou průzkumové analýzy dat vyšetřete, zda jsou splněny požadavky, kladené na náhodný výběr a zda je splněn i požadavek čistoty kalciferolu.

(2) Vyčíslete také kvantilové charakteristiky šikmosti a špičatosti: polosumu *ZL*, rozpětí *RL*, šikmost *SL*, pseudosigmu *GL* a délky konců *TL* pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. *ZL* a *SL*), rozptýlení (tj. *RL*) a špičatost (tj. *GL* a *TL*).

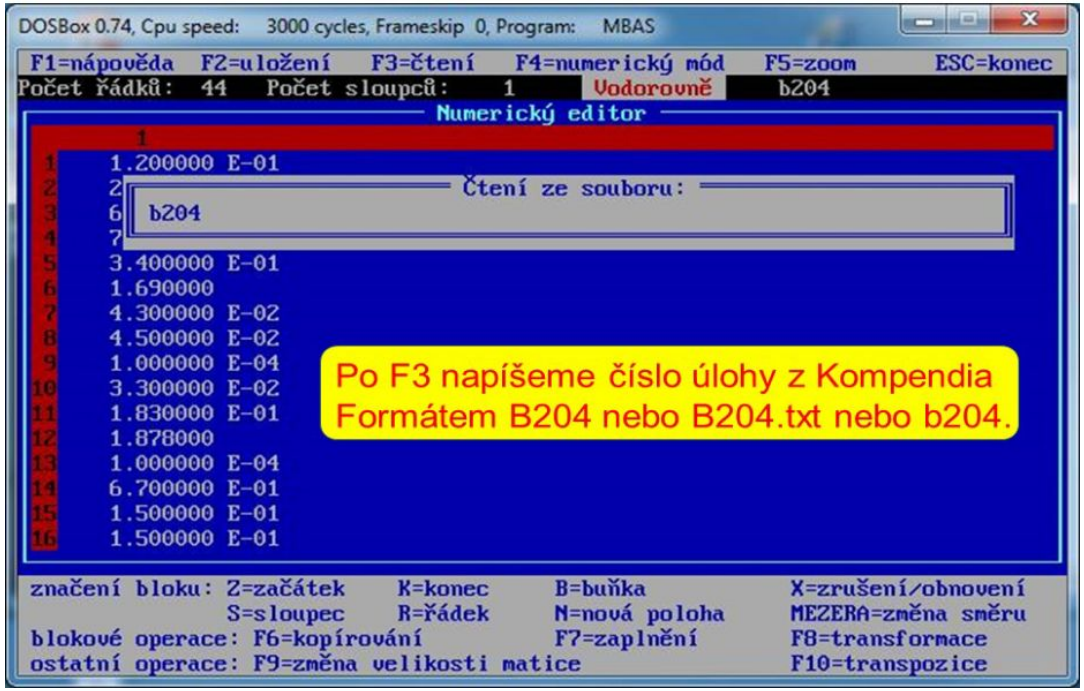
(3) Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti, (symboly viz [19]).

(4) Které diagnostiky shodně indikují vybočující hodnoty?

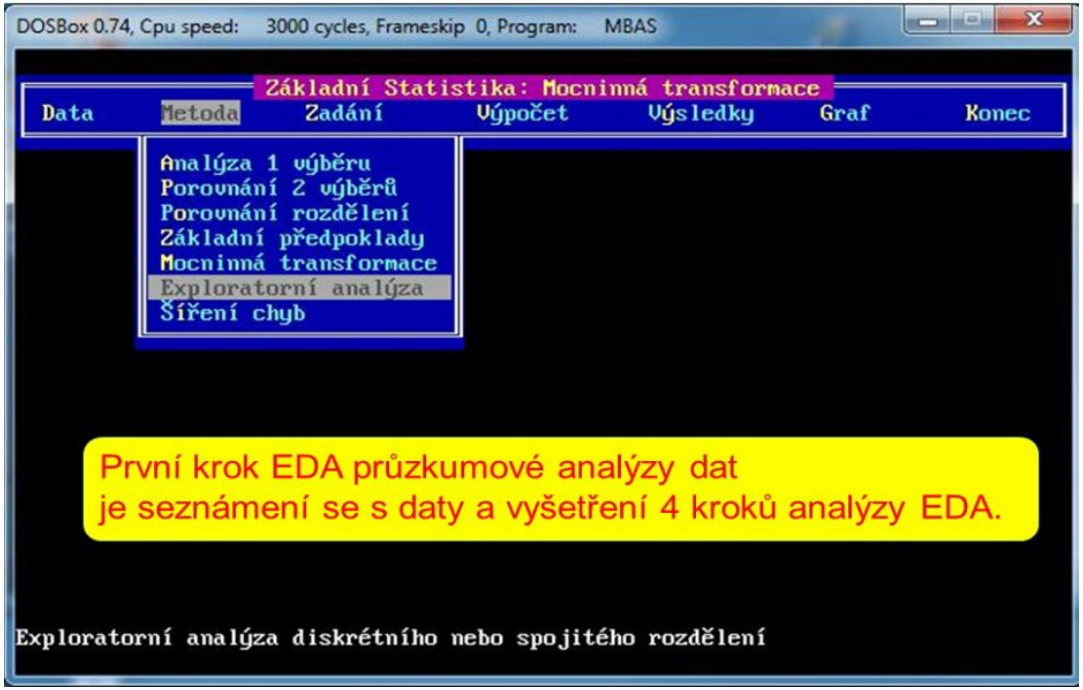
(5) Jak velké procento hodnot dosahuje nižšího obsahu než 0.4 %?

Data: Obsah ergosterinu [%]

Zadání dat: formou kódovaného čísla úlohy z Kompendia

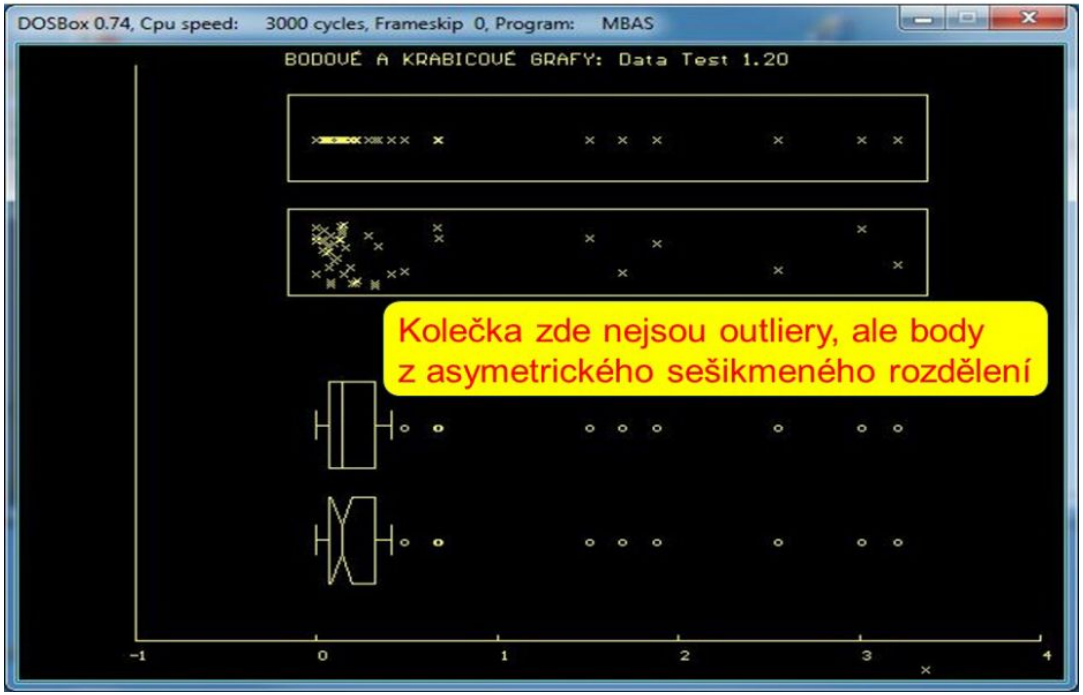


Volba metody: Exploratorní analýza dat

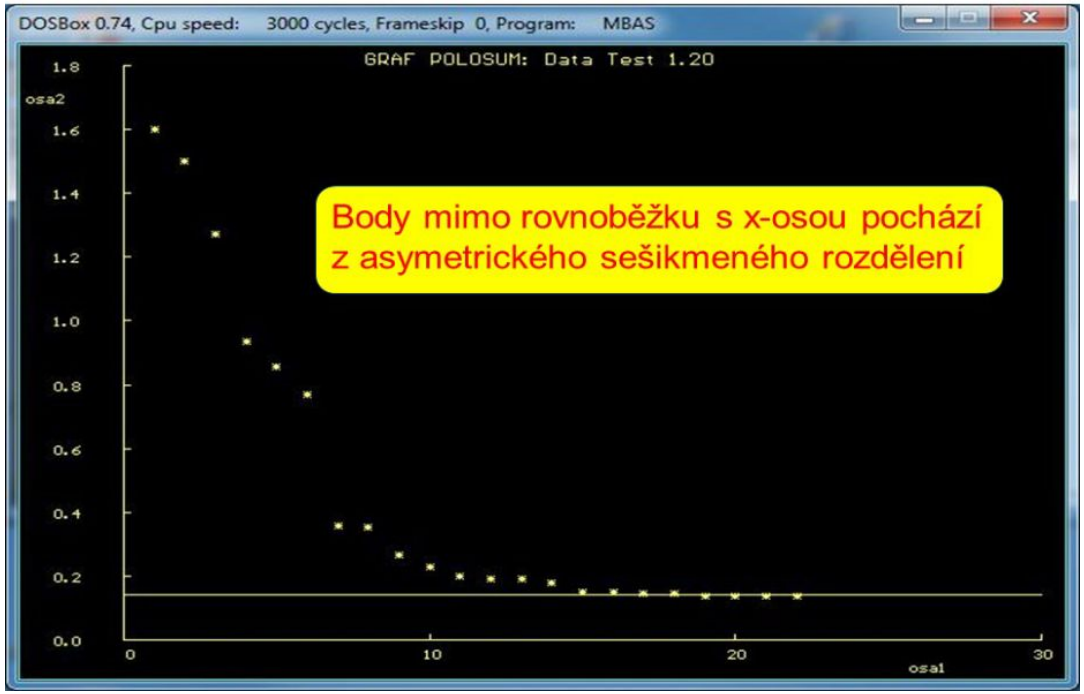


Diagnostikování: Kvantilový graf ukazuje na dlouhý ocas a sešikmené rozdělení

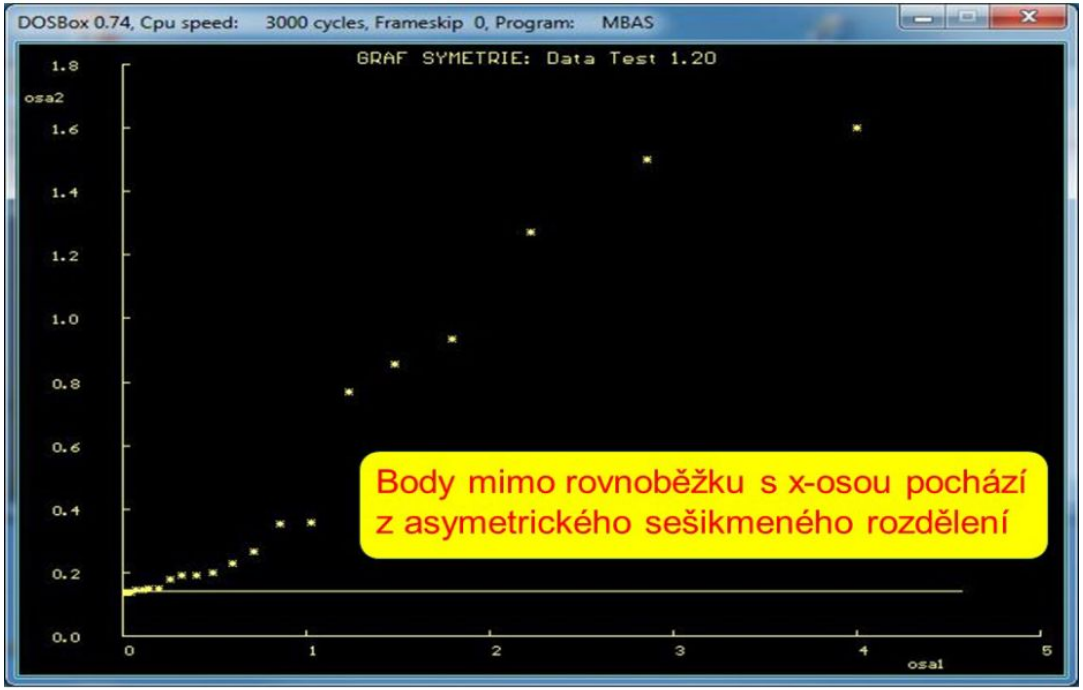
Diagnostikování: Bodové grafy ukazují řadu odlehlých bodů nahoře



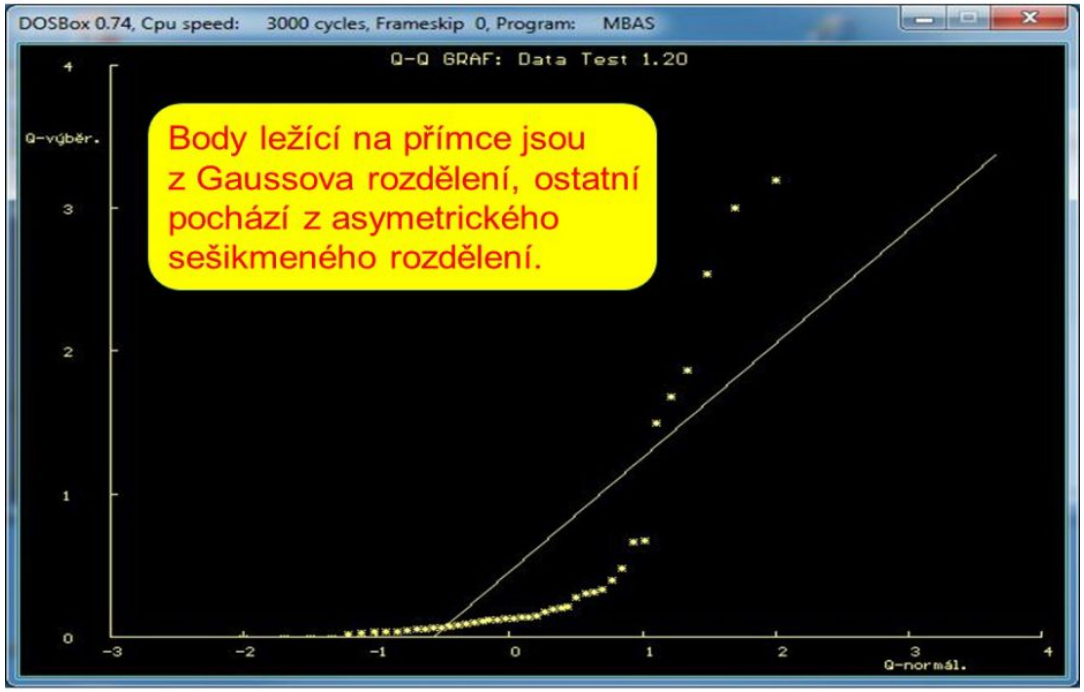
Diagnostikování: Graf polosum odhaluje silně sešikmené, asymetrické rozdělení



Diagnostikování: Graf symetrie odhaluje silně sešikmené, asymetrické rozdělení



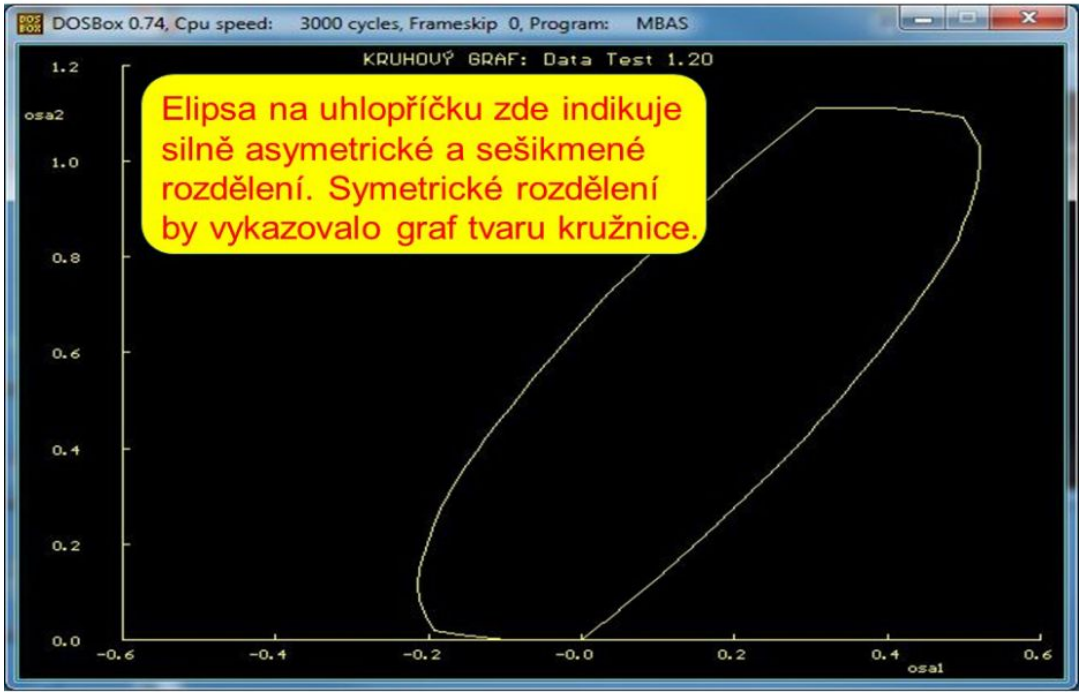
Diagnostikování: Q-Q graf odhaluje silně sešikmené, asymetrické rozdělení



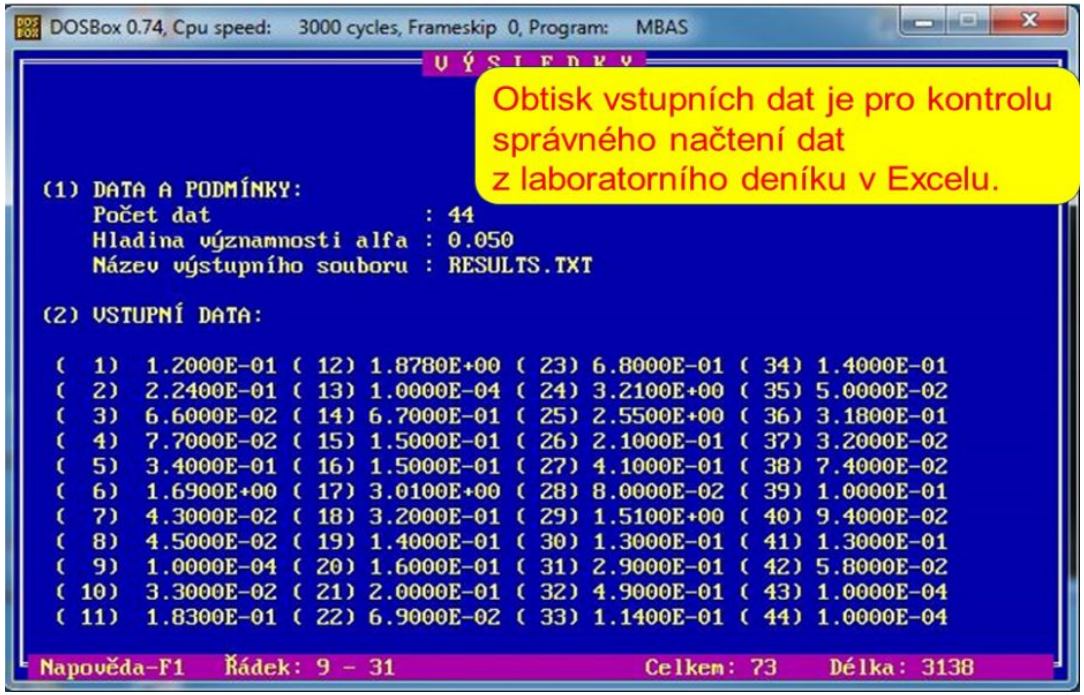
Diagnostikování: Graf hustoty pravděpodobnosti ukazuje na špičaté asymetrické r.

Diagnostikování: Graf rozptýlení s kvantily ukazuje na sešikmené, asymetrické r.

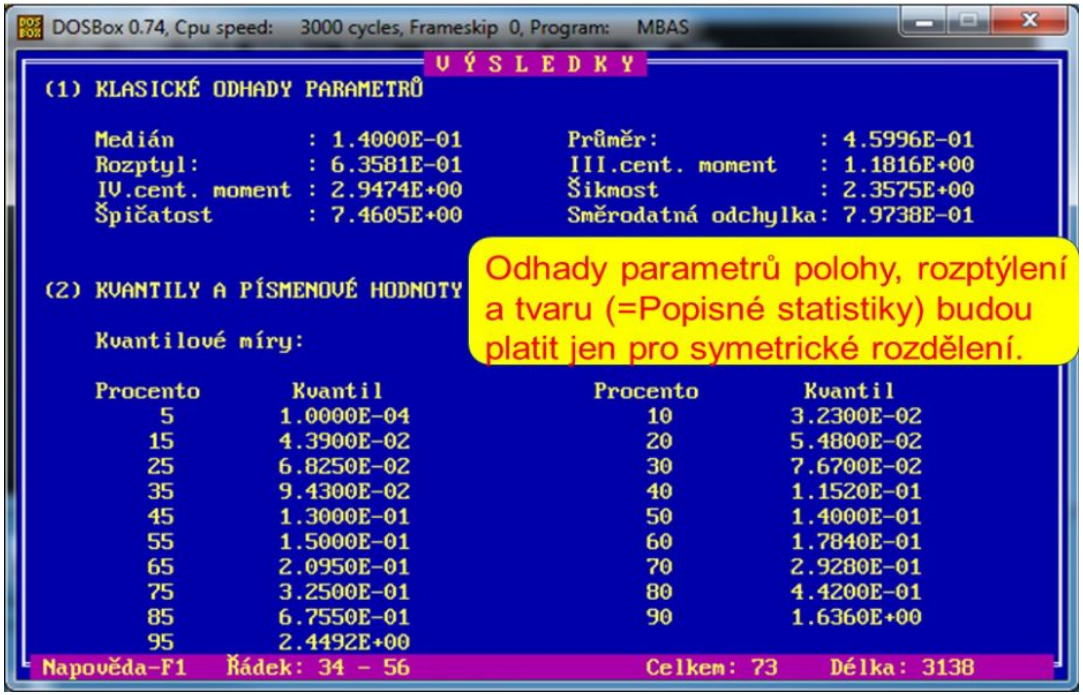
Diagnostikování: Kruhový graf ukazuje na sešikmené, asymetrické rozdělení



Output: Kontrola vstupních dat, které jdou do analýzy EDA



Output: Klasické odhady a robustní kvantilové odhady parametrů



Odhady parametrů polohy, rozptýlení a tvaru (=Popisné statistiky) budou platit jen pro symetrické rozdělení.

EDA průzkumová analýza dat vede ve svých diagnostikách k prvnímu závěru o datech, který je třeba zapsat do protokolu.

Závěr EDA:

Rozdělení je silně asymetrické a odlehlé hodnoty nelze odstranit jako fatální hodnoty, protože tyto rovněž patří do silně sešikmeného rozdělení. Odstraněním bodů bychom ztratili informaci. Vhodnou metodou k odhadu parametrů polohy, rozptýlení bude proto jedině některá z transformací původních dat.

Output: Písmenové hodnoty a kvantilové míry tří parametrů

Výhodné kvantilové míry nejsou citlivé na tvar rozdělení dat, a proto nám správně odhadnou o datech vše co potřebujeme.

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

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

(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.

Test významnosti autokorelačního koeficientu ρ_a:
Hypotéza: nulová H_0: $\rho_a = 0$, a alternativní H_A: $\rho_a \neq 0$.
Testační kritérium:

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

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.

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.

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 $8 \le n \le 100$ použít aproximace

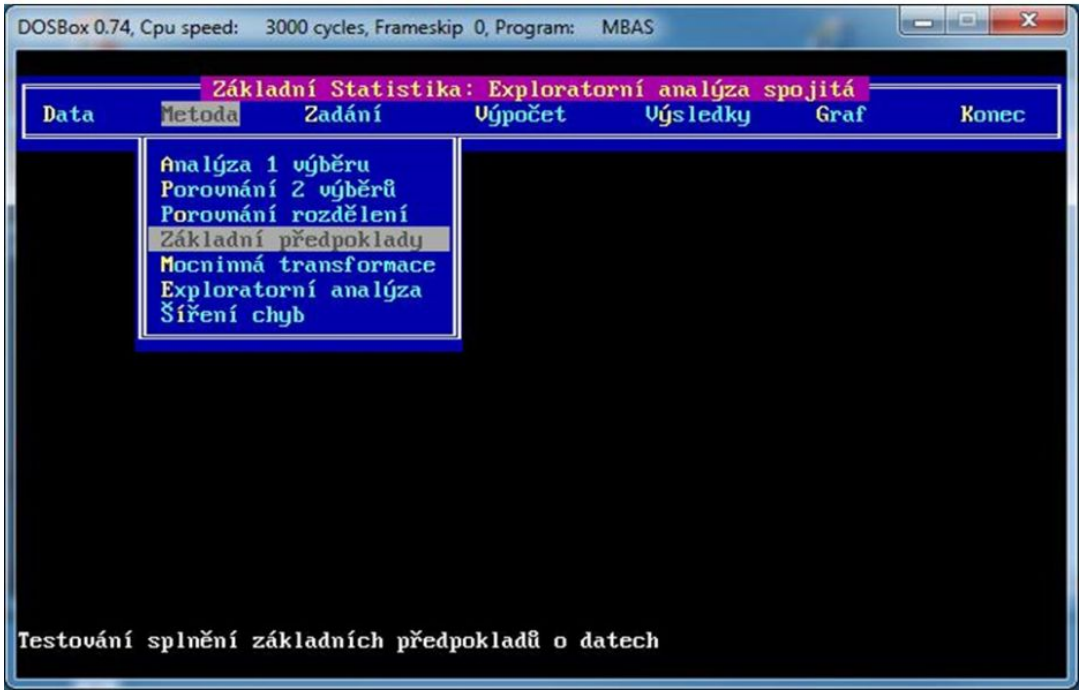
$$K \approx 2.25 - \frac{3.6}{n}$$

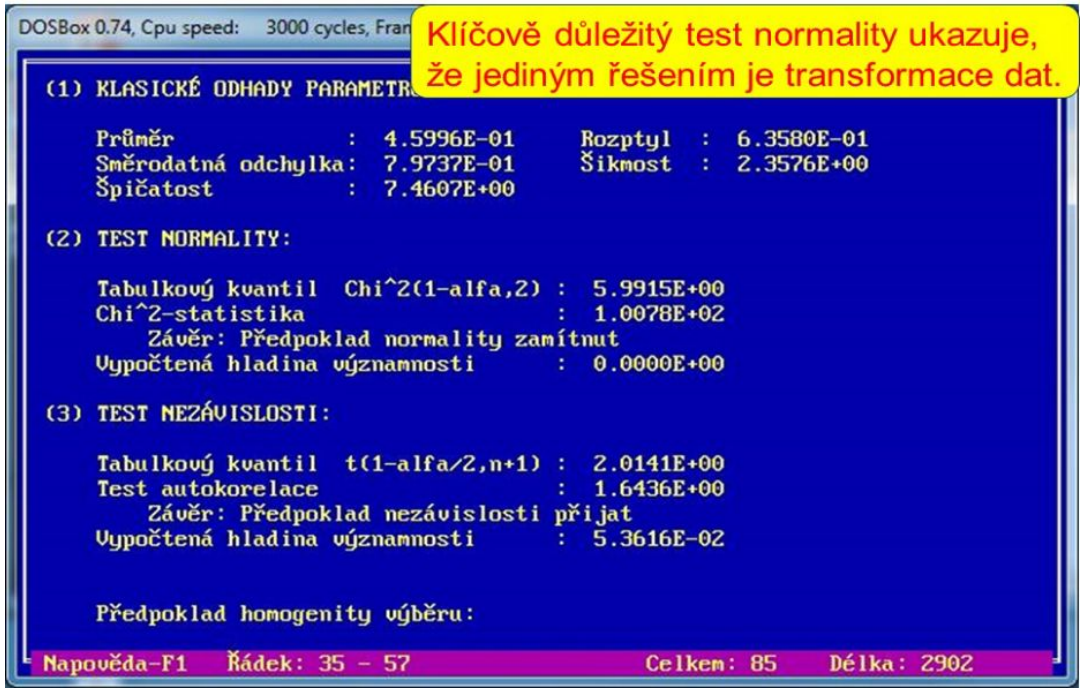
Pro takto určený parametr K se všechny prvky výběru ležící mimo hradby $[B_D^*, B_H^*]$ považují za vybočující.

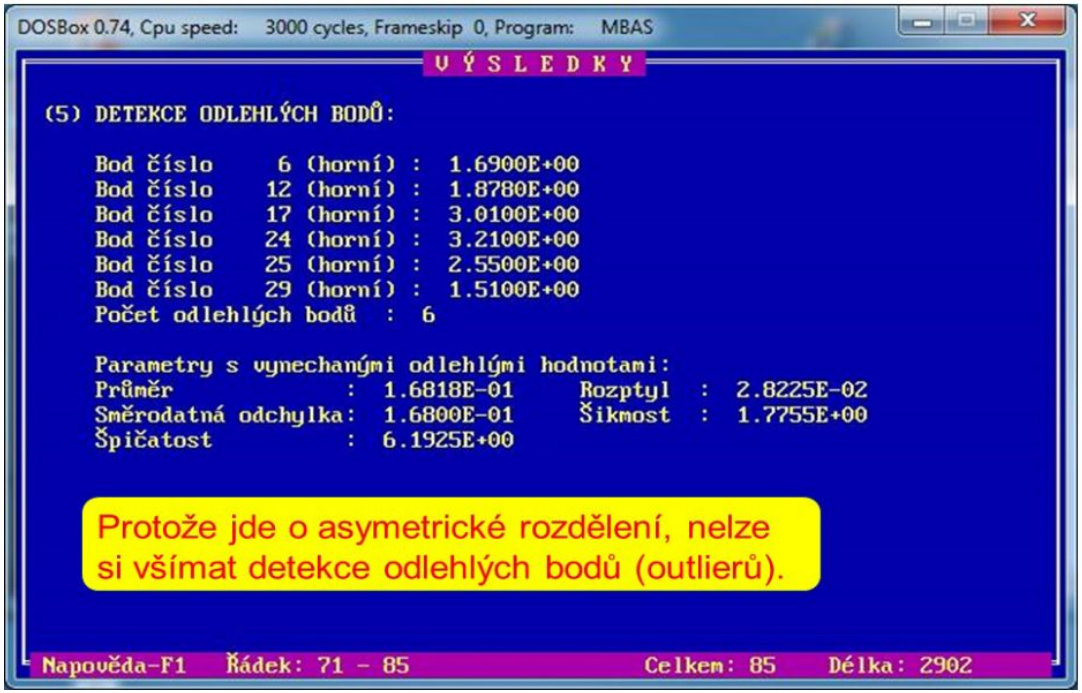
2. kapitola
EDA: Ověření předpokladů o výběru

Řešení úloh z Kompendia:

B204







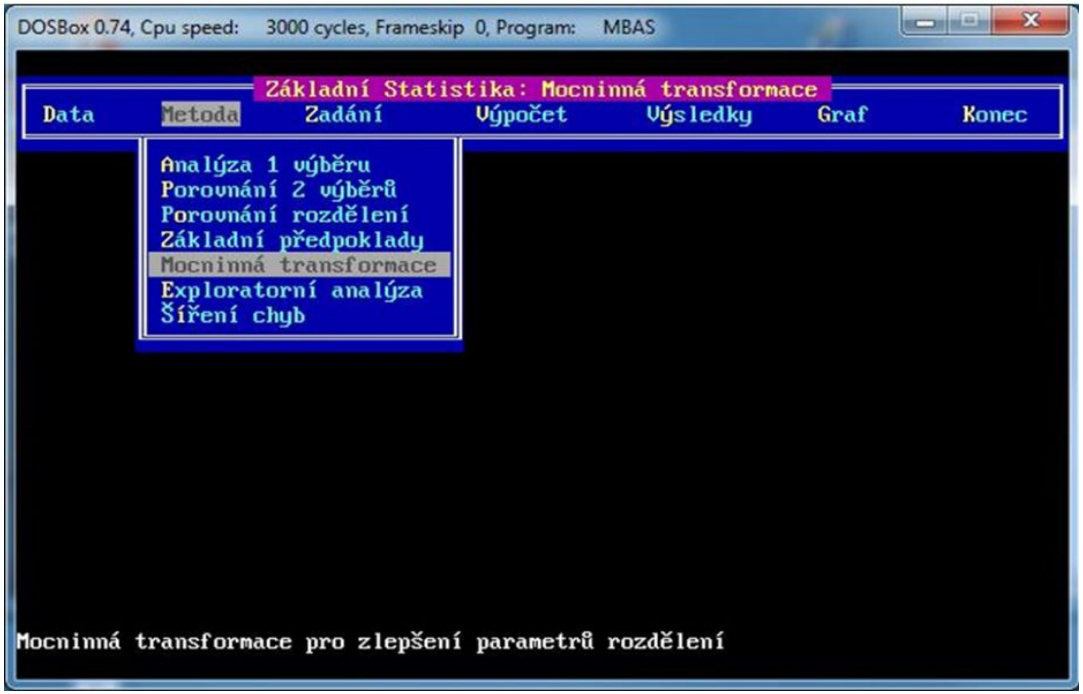
2. kapitola

EDA: Transformace dat

Řešení úloh z Kompendia:

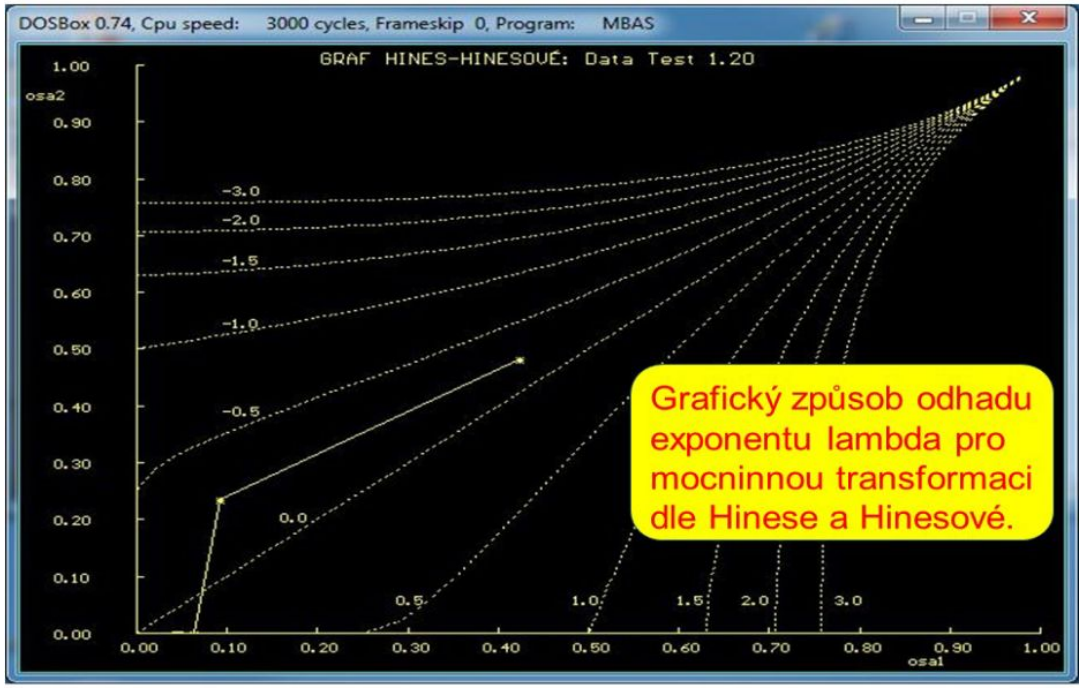
B204

Volba metody: Mocninná transformace a Box-Coxova transformace

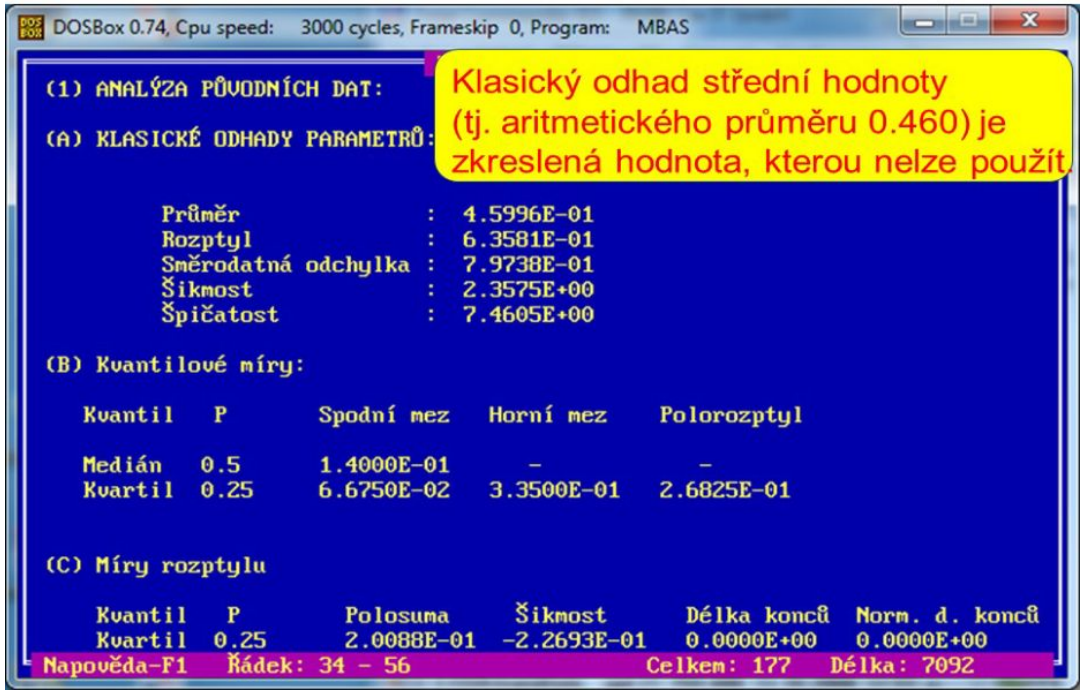


Diagnostika: Rozhodčí kritérium k užití transformace a odhad mocniny

Diagnostika: Graf Hinesové a Hinese k určení mocniny lambda



Output: Klasické odhady původních dat



Output: Opravený průměr po prosté mocninné transformaci

```
DOSBox 0.74, Cpu speed:   3000 cycles, Frameskip  0, Program:   MBAS
                          V Ý S L E D K Y
(2) PROSTÁ MOCNINNÁ TRANSFORMACE:

(A) Optimální hodnoty mocniny pro vybraná kritéria:

    Optimální mocnina: 1.3333E-01 pro šikmost          : 5.0711E-01
    Optimální mocnina: 2.6667E-01 pro špičatost        : 3.4145E+00
    Optimální mocnina: 1.3333E-01 pro asymetrii        : 4.9216E-03
    Optimální mocnina:-2.6667E-01 pro asymetrii, rob.  : 2.5736E-02
    Optimální mocnina: 1.3333E-01 pro Hinkley-asymetrii: 3.6806E-03

    Zvolená mocnina  : 0.13

         Průměr             :  7.7371E-01
         Rozptyl            :  4.3357E-02
         Směrodatná odchylka :  2.0822E-01
         Šikmost            : -5.0711E-01
         Špičatost          :  3.8571E+00
         Opravený průměr    :  1.4599E-01

(B) Kvantilové míry:

   Kvantil   P

 Napověda-F1   Řádek:
```

U mocninné transformace použijeme opravený (tj. retransformovaný) průměr 0.146, který je rigorózním nejlepším odhadem.

Output: Opravený průměr po Box-Coxově transformaci

```
DOSBox 0.74, Cpu speed:   3000 cycles, Frameskip  0, Program:   MBAS
                          V Ý S L E D K Y
(3) BOX-COXOVA TRANSFORMACE:

(A) Optimální hodnoty mocniny pro vybraná kritéria:

    Optimální mocnina: 1.3333E-01 pro šikmost          : 5.0711E-01
    Optimální mocnina: 2.6667E-01 pro špičatost        : 3.4145E+00
    Optimální mocnina: 1.3333E-01 pro asymetrii        : 4.9216E-03
    Optimální mocnina:-2.6667E-01 pro asymetrii, rob.  : 2.5736E-02
    Optimální mocnina: 1.3333E-01 pro Hinkley-asymetrii: 2.7604E-02
    Optimální mocnina: 1.3333E-01 pro věrohodnost      : 6.7101E+01

    Zvolená mocnina  : 0.13

         Průměr             : -1.6972E+00
         Rozptyl            :  2.4388E+00
         Směrodatná odchylka :  1.5617E+00
         Šikmost            : -5.0711E-01
         Špičatost          :  3.8571E+00
         Opravený průměr    :  1.4599E-01

(B) Kvantilové míry:

 Napověda-F1   Řádek:
```

U Box-Coxovy transformace použijeme opravený (tj. retransformovaný) průměr 0.146, který je rigorózním nejlepším odhadem.

Output: Setříděná původní a transformovaná data

Ukázka původních a transformovaných dat po setřídění.

```
DOSBox 0.74, Cpu speed:   3000 cycles, Frameskip  0, Program:   MBAS
                          V Ý S L E D K Y
(4) SETŘÍDĚNÁ PŮVODNÍ A TRANSFORMOVANÁ DATA:

     Původní        Po prosté transformaci  Po Box-Coxově-transformaci

    1.0000E-04          2.9286E-01              -5.3035E+00
    1.0000E-04          2.9286E-01              -5.3035E+00
    1.0000E-04          2.9286E-01              -5.3035E+00
    1.0000E-04          2.9286E-01              -5.3035E+00
    3.2000E-02          6.3196E-01              -2.7603E+00
    3.3000E-02          6.3455E-01              -2.7408E+00
    4.3000E-02          6.5735E-01              -2.5699E+00
    4.5000E-02          6.6135E-01              -2.5399E+00
    5.0000E-02          6.7070E-01              -2.4697E+00
    5.8000E-02          6.8411E-01              -2.3692E+00
    6.6000E-02          6.9599E-01              -2.2800E+00
    6.9000E-02          7.0013E-01              -2.2490E+00
    7.4000E-02          7.0669E-01              -2.1998E+00
    7.7000E-02          7.1045E-01              -2.1716E+00
    8.0000E-02          7.1408E-01              -2.1444E+00
    9.4000E-02          7.2960E-01              -2.0280E+00
    1.0000E-01          7.3564E-01              -1.9827E+00
    1.1400E-01          7.4861E-01              -1.8854E+00
    1.2000E-01          7.5374E-01              -1.8469E+00
 Napověda-F1   Řádek: 130 - 152          Celkem: 177   Délka: 7092
```

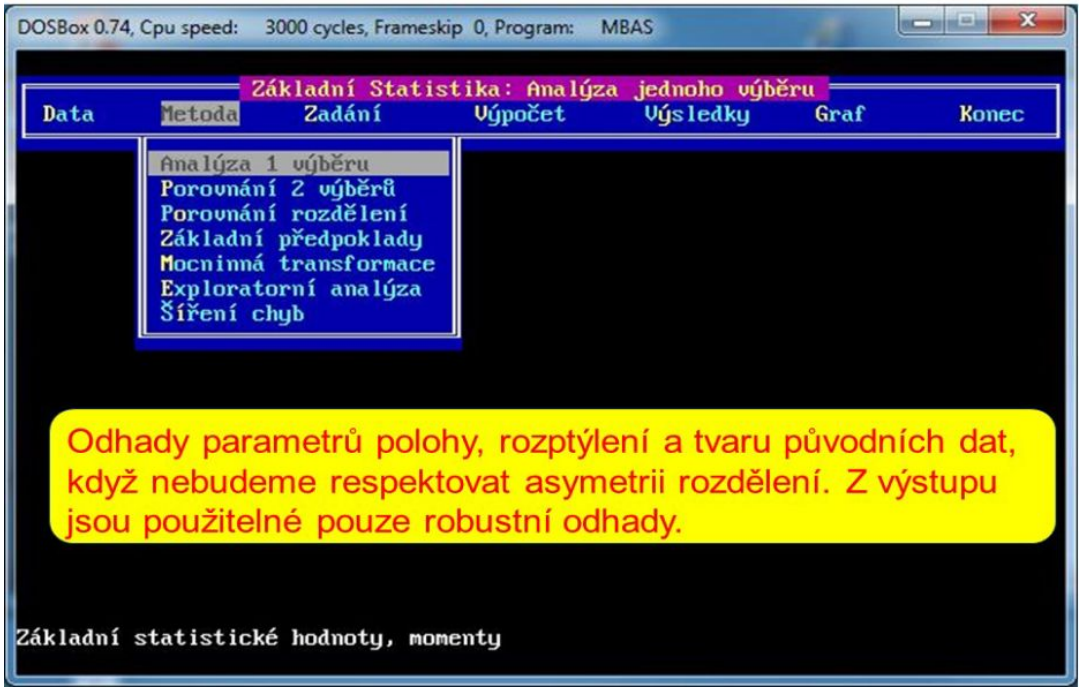
2. kapitola

EDA: Odhady UDA

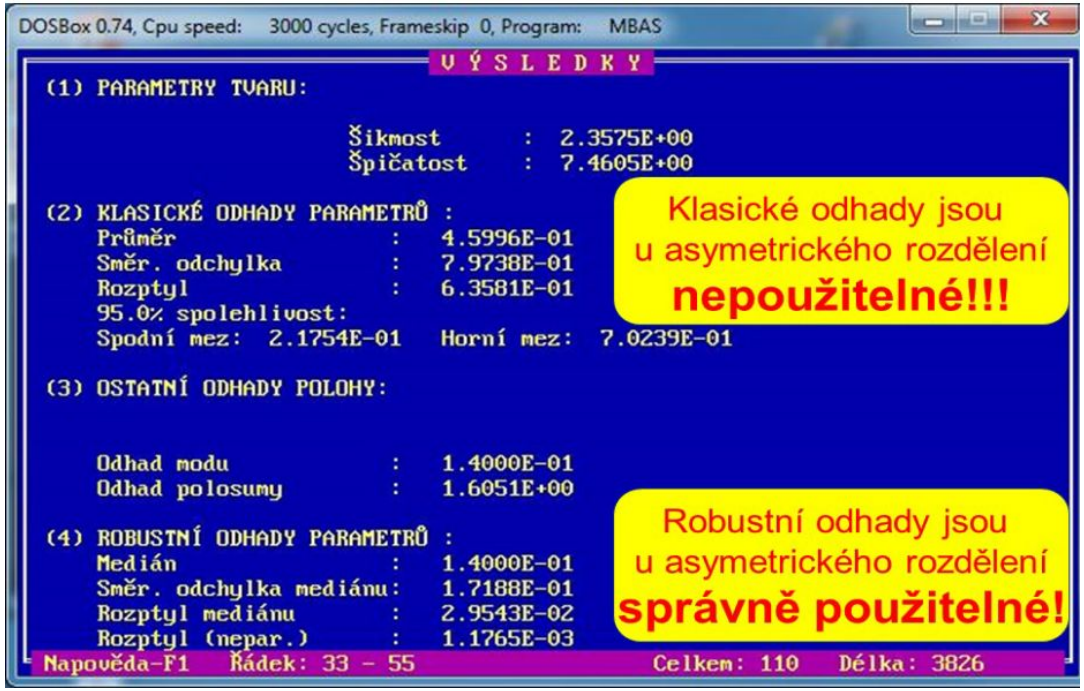
Řešení úloh z Kompendia:

B204

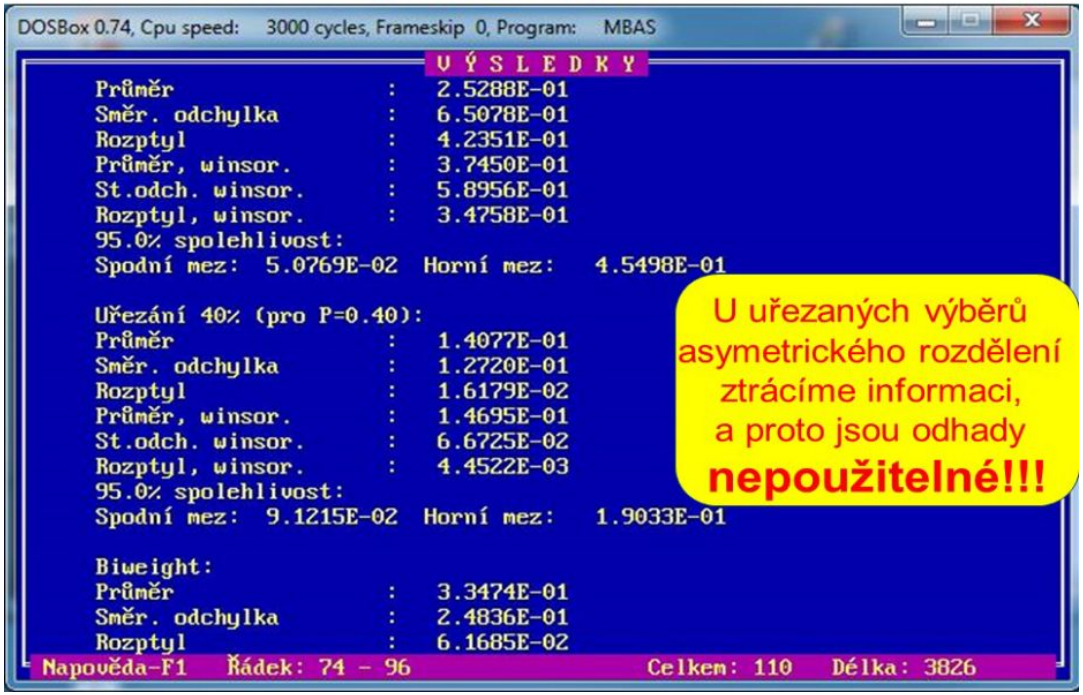
Volba metody



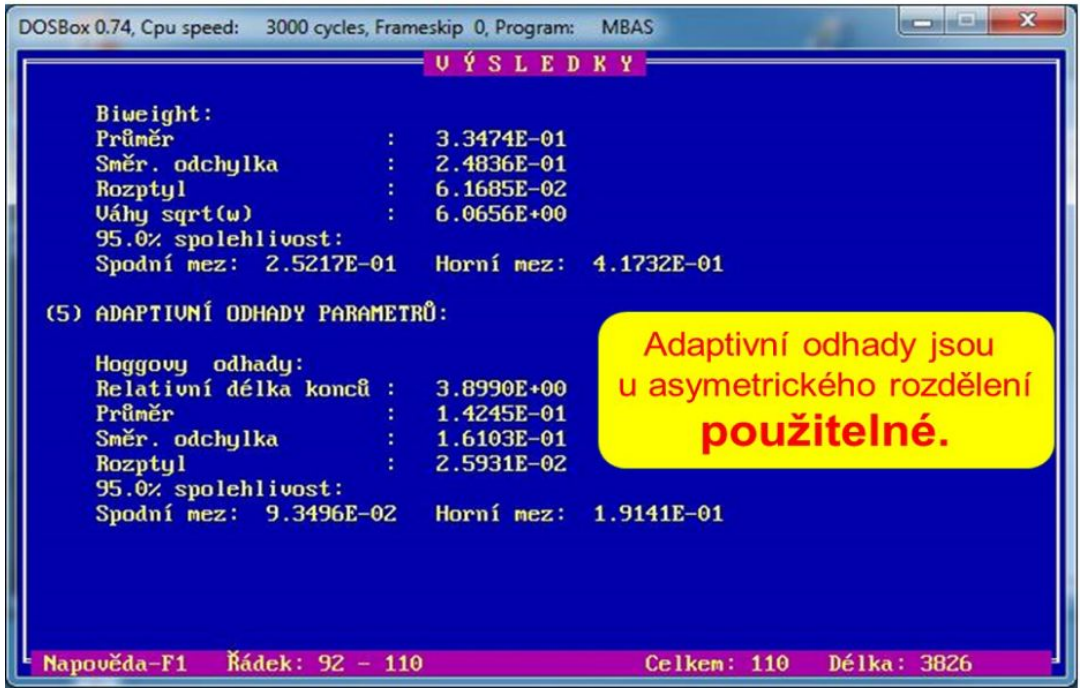
Output: Klasické a robustní odhady polohy, rozptýlení a tvaru



Output: Robustní odhady polohy, rozptýlení a tvaru



Output: Adaptivní odhady polohy, rozptýlení a tvaru



EDA vede ve svých diagnostikách k jedinému závěru o správných odhadech metodou transformace dat, což je třeba zapsat do protokolu.

Závěr: Mezi nejlepší odhady polohy, rozptýlení a tvaru (bodové a intervalové) patří retransformované odhady po mocninné a Box-Coxově transformaci.

2. kapitola

EDA: Exploratorní analýza dat

Řešení úloh z Kompendia:

B201

2. kapitola

Porovnání programů ADSTAT a QCEXPERT

Při řešení úloh z Kompendia

Úloha B2.01 *Rozdělení obsahu léčiva v krvi u náhodně vybraných pacientů*

Byl sledován obsah léčiva v krvi u náhodně vybraných pacientů.

(1) Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru a rozhodněte o typu rozdělení.

(2) Vyšetřete předpoklady o náhodnosti a normalitě výběru a sestrojte histogram.

(3) Obsahují data za předpokladu, že pocházejí z normálního rozdělení, nějaké odlehlé body?

(4) Jaký maximální obsah léčiva v krvi má 75% pacientů?

(5) Odhadněte hloubku prvku 4.00?

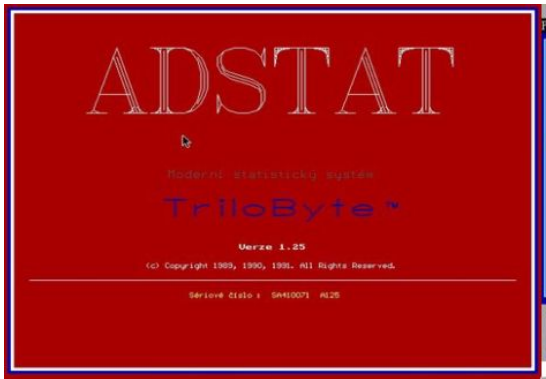
Data: Obsah léčiva v krvi [mg.l!1]

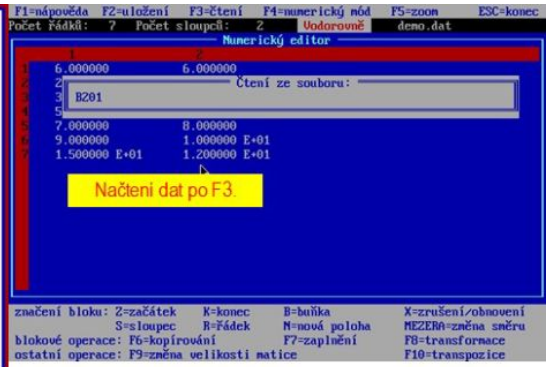
3.86 4.06 3.67 3.97 3.76 3.61 3.76 4.26 3.52 3.96

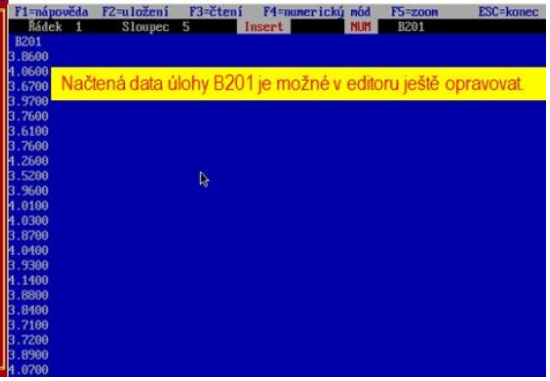
.. ..

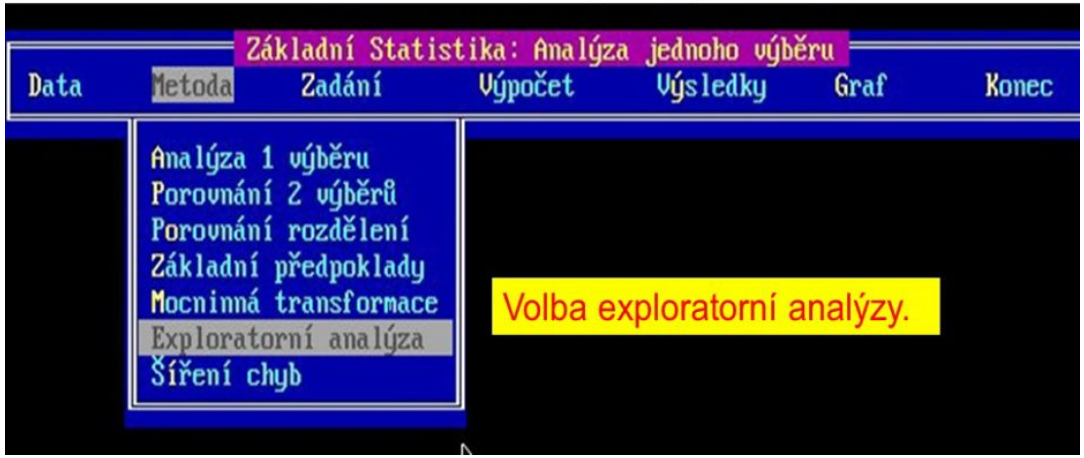
1. Průzkumová analýza dat:

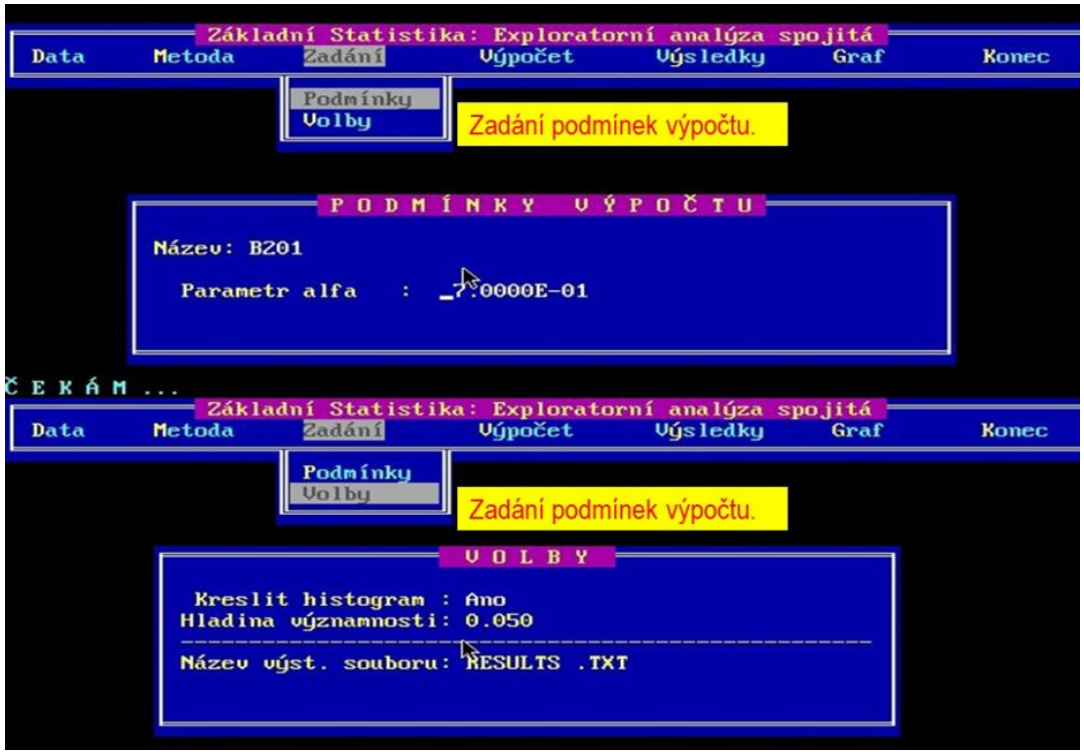
Diagnostické grafy: *stupeň symetrie rozdělení*
lokální koncentrace dat
vybočující data

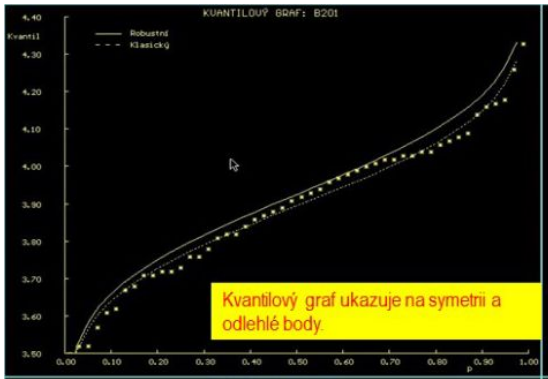


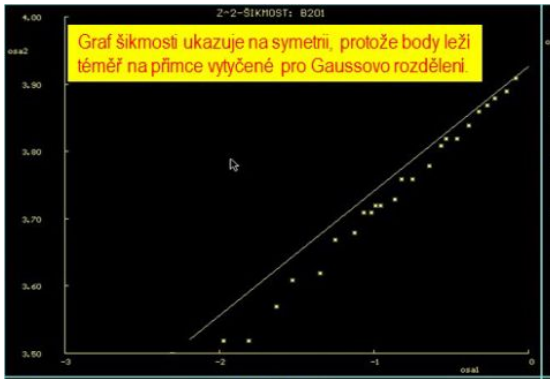


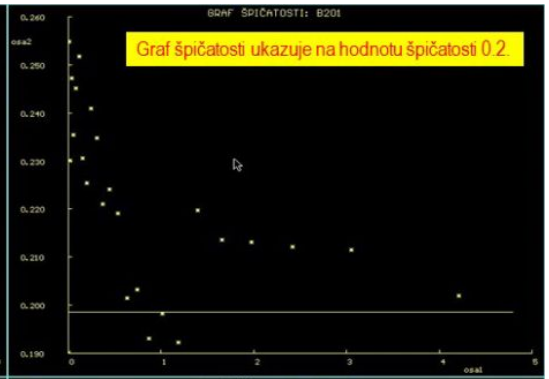


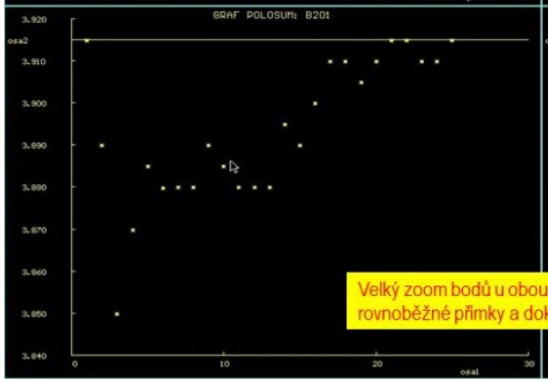


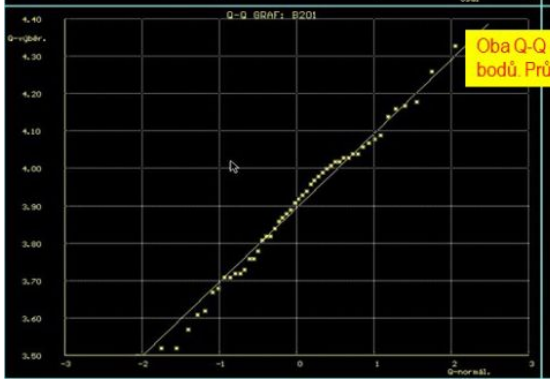


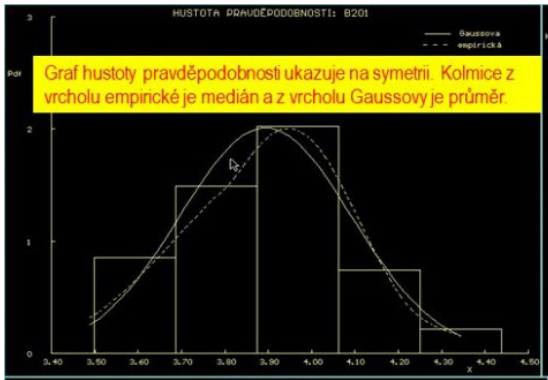


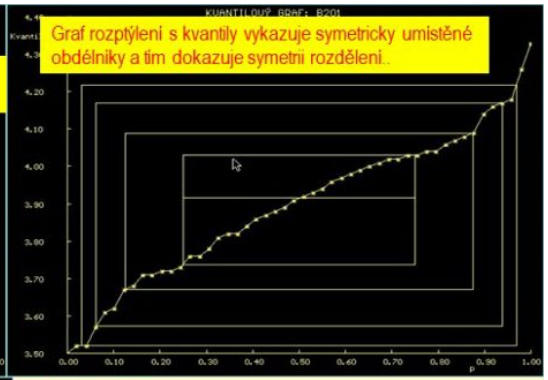










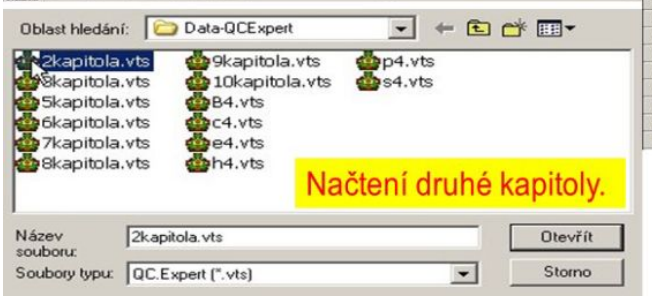


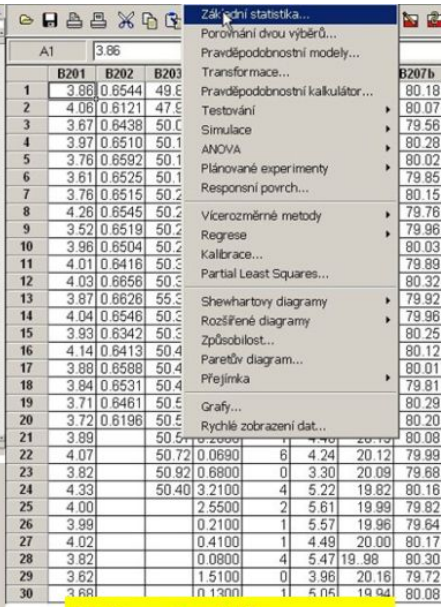
Závěr EDA:

1. Rozdělení je symetrické (mezi rovnoměrným a Gaussovým).

2. Rozdělení obsahuje nahoře 2 a dole cca 5 odlehlých bodů – outlierů.

3. Není asi nutné užít transformaci dat.





Data a parametry: Data jsou organizována do sloupců. V prvním řádku jsou názvy sloupce. Jsou-li ve vstupním dialogovém panelu vybrány "Všechna data" nebo "Sloupce", mohou být délky jednotlivých sloupců různé. Má-li se výpočet provést pro "Průměry podskupin, měly by všechny slupce mit stejnou délku. Minimální počet sloupců je 1. Minimální počet dat je 2.

Řád trendu určuje z kolika po sobě jdoucich dat budou počitány klouzavé průměry a klouzavé mediány Hodnota by měla být menši než polovina počtu dat.

Testuj hodnotu zadává se hodnota pro T-test. Program testuje na zadané hladině významnosti, zda tato hodnota může být shodná se střední hodnotou dat.

Vyhlazení hustoty udává šířku jádra pro jádrové vyhlazení pro graf hustoty pravděpodobnosti. Čím je větší, tím bude křivka hustoty pravděpodobnosti hladší a naopak. Hodnota musí být větší než nula.

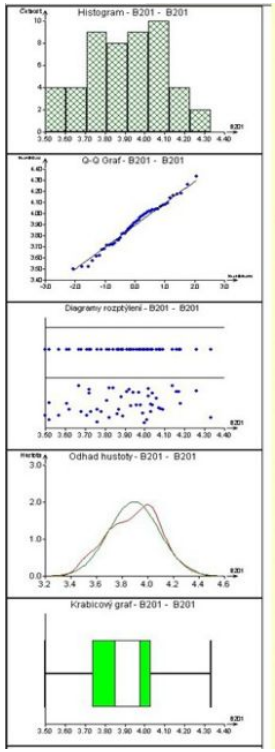
Řád autokorelace udává do kterého řádu se budou počítat autokorelační koeficienty. Hodnota musí být alespoň o 2 menší než počet platných dat.

Hladina významnosti udává spolehlivost pro intervaly spolehlivosti a statistické testy. Musí být větší než nula a menší než 0.5. Vynásobena 100 udává hodnotu v procentech. Obvyklá hodnota je 0.05 (tedy 5%).

Všechna data všechny vybrané sloupce budou brány jako jediný sloupec.

Sloupce Výpočet se provede pro každý z vybraných sloupců zvlášť.

Průměry podskupin Výpočet se provede pro řádkové průměry z vybraných sloupců. Pokud nejsou sloupce stejně dlouhé, nebo obsahují chybějící data, provede se výpočet jen pro úplné řádky. Tento výpočet má význam především pro diagnostiku dat pro regulační diagramy typu x-průměr.



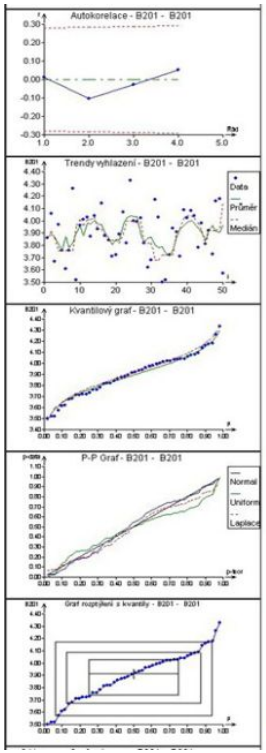
Grafy výstupu Exploratorní analýzy a jejich vysvětlení.

Histogram: nejstarší diagram rozdělení ukazuje na symetrii rozdělení. Histogram četnosti dat v jednotlivých třídách s konstantní šířkou; optimální počet tříd je stanovován automaticky s ohledem na počet dat.

Q-Q graf: kvantil-kvantilový graf vykazuje většinu bodů na přímce.
4 body dole a 2 body nahoře jsou podezřelé, že jsou odlehlé. Graf pro diagnostiku normality a odlehlých měření; pro normální data bez odlehlých měření má tvar přímky; pro normální data s odlehlými měřeními má tvar přímky s koncovými body ležícími mimo tuto přímku; pro systematicky sešikmená data s kladnou šikmostí (např. rozdělení lognormální, exponenciální) má nelineární konvexní tvar . Pro systematicky sešikmená data se zápornou šikmostí má nelineární konkávní tvar .

Diagramy rozptýlení: horní diagram rozptýlení a dolní rozmítnutý diagram rozptýlení ukazují na podezřelé body z odlehlosti. Zobrazuje všechna data ve skutečném měřítku na ose X. Popis osy Y nemá význam. Aby nedošlo ke splývání shodných nebo blízkých dat, jsou ve spodní polovině grafu zobrazena táž data, ale náhodně rozmítnutá ("rozházená") na ose Y.

Jádrový odhad hustoty pravděpodobnosti: zelená křivka značí proložení teoretickým Gaussovým normálním rozdělením na bázi aritmetického průměru a směrodatné odchylky, zatímco červená křivka značí robustní proložení skutečným experimentálním rozdělením na bázi mediánu a kvartilového rozpětí. Porovnání průběhu hustoty pravděpodobnosti normálního rozdělení (plná



Graf autokorelace: data jsou nezávislá, protože leží mezi červenými intervaly.

Graf autokorelačních koeficientů až do řádu autokorelace zadaného v dialogovém panelu,. Červené meze ohraničují interval, v němž jsou koeficienty statisticky nevýznamné na zadané hladině významnosti. Překročí-li některý autokorelační koeficient tyto meze, je třeba považovat za závislá.

Graf trendů vyhlazení: v datech není trend.

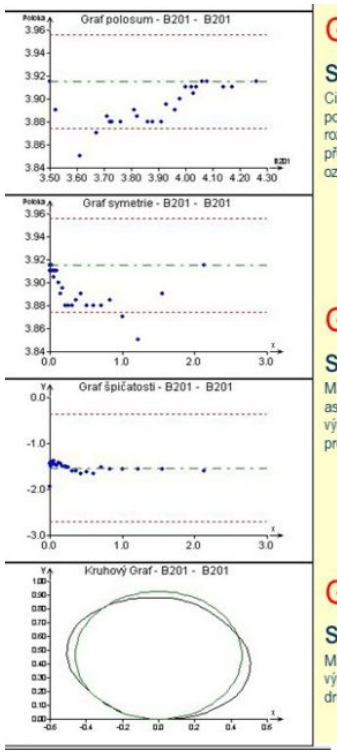
Grafické znázornění trendu v datech pomocí klouzavého průměru (plná křivka) a klouzavého mediánu (přerušovaná křivka) na základě "Řádu trendu" zadaného v dialogovém panelu. Čím je řád trendu větší, tím jsou křivky hladší, méně citlivé na lokální poruchy a zachycují spíše globální dlouhodobý trend. Menší řád zachycuje spíše lokální chování dat. Klouzavý medián je méně citlivý (robustní) na lokální poruchy v datech a jednotlivá odlehlá měření a tedy vhodnější pro tyto případy.

Kvantilový graf: ukazuje na Gaussovo rozdělení.

Zobrazuje empirické kvantily dat proložené kvantilovou funkcí normálního rozdělení. Zelená křivka odpovídá funkci s klasickým průměrem a rozptylem (nerobustní), červená křivka odpovídá mediánu a mediánové odchylce (robustní). Podle toho, která z křivek lépe prokládá data, je vhodné zvolit jako odhad střední hodnoty průměr nebo medián.

P-P graf: podobnost exp. křivky s normální ukazuje na Gaussovo rozdělení.

Porovnává data s normálním (modrá křivka plná), Laplaceovým (zelená křivka) a rovnoměrným rozdělením pomocí teoretické a empirické distribuční funkce. Která křivka leží nejblíže černé přímce $y = x$, to rozdělení odpovídá experimentálním datům. Graf slouží pro rozlišení symetrických rozdělení podle špičatosti. Podobnost rovnoměrnému rozdělení ukazuje na možné vyloučení vysokých a nízkých hodnot, podobnost s Laplaceovým rozdělením ukazuje na možnou nekonstantnost rozptylu dat.



Graf polosum: většina bodů je v mezích a proto jde o symetrické rozdělení.

Citlivý indikátor asymetrie rozdělení. V ideálním případě leží body na horizontální přímce. Horizontální přímka, na niž leží poslední bod, představuje medián a červené přerušované meze jeho interval spolehlivosti. V případě asymetrického rozdělení vykazují body výrazný trend (rostoucí pro zápornou šikmost, nebo klesající, pro kladnou šikmost) výrazně překračující přerušované meze. Body jsou konstruovány ze dvojic dat (první, poslední; druhý, předposlední; atd.), proto označením bodu jsou označena dvě příslušná data v tabulce.

Graf symetrie: většina bodů je v mezích a proto jde o symetrické rozdělení.

Má podobný význam jako předchozí graf polosum. Směrnice případného trendu je úměrná šikmosti. V případě asymetrického rozdělení vykazují body výrazný trend (rostoucí pro zápornou šikmost, nebo klesající, pro kladnou šikmost) výrazně překračující přerušované meze. Body jsou konstruovány ze dvojic dat (první, poslední; druhý, předposlední; atd.), proto označením bodu jsou označena dvě příslušná data v tabulce.

Graf špičatosti: většina bodů je v mezích a proto jde o symetrické rozdělení.

Má podobný význam jako dva předchozí grafy. Směrnice případného trendu je úměrná odchylky špičatosti od 3. V případě výrazně nenormální špičatosti rozdělení vykazují body výrazný trend. Body jsou konstruovány ze dvojic dat (první, poslední; druhý, předposlední; atd.), proto označením bodu jsou označena dvě příslušná data v tabulce.

Základní analýza dat

Název úlohy :	B201
Data:	Všechna
Řád trendu :	4
Testovaná hodnota :	0
Vyhlazení hustoty :	0.5
Hladina významnosti :	0.05
Název sloupce :	B201
Počet platných dat :	50

Sloupec Název sloupce.
Počet řádků Celkový počet řádků v datech.
Počet platných dat Celkový počet platných dat v datech.
Počet chybějících hodnot Počet prázdných buněk v datech.

Klasické parametry :

Název sloupce :	B201
Průměr :	3.894
Spodní mez :	3.837588535
Horní mez :	3.950411465
Rozptyl :	0.0394
Směr. odchylka :	0.1984943324
Šikmost	-0.119116936
Odchylka od 0 :	Nevýznamná
Špičatost :	2.40159769
Odchylka od 3 :	Nevýznamná
Polosuma	3.915
Modus :	3.955352941

Klasické parametry

Aritmetický průměr Odhad střední hodnoty pro normálně rozdělená data.
Spodní mez Spodní mez intervalu spolehlivosti aritmetického průměru na zadané hladině významnosti.
Horní mez Horní mez intervalu spolehlivosti aritmetického průměru na zadané hladině významnosti.
Rozptyl Odhad rozptylu.
Směrodatná odchylka Druhá odmocnina z rozptylu.
Šikmost Odhad třetího statistického momentu, šikmosti.
Rozdíl od 0 Normální a každé symetrické rozdělení má šikmost nulovou. Je-li hodnota šikmosti statisticky významně odlišná od 0, nelze data považovat za symetrická. Spolehlivější je však test normality.
Špičatost Odhad čtvrtého statistického momentu, špičatosti.
Rozdíl od 3 Normální rozdělení má špičatost 3. Je-li hodnota špičatosti statisticky významně odlišná od 3, lze předpokládat, že data neodpovídají normálnímu rozdělení. Spolehlivější je však test normality.
Polosuma Odhad polosumy, tedy středu nejmenší a největší hodnoty.
Modus Odhad modu rozdělení, tedy maxima na křivce hustoty pravděpodobnosti.

Robustní parametry :

Název sloupce :	B201
Medián :	3.915
IS spodní :	3.709937536
IS horní :	4.120062464
Medianová směr. odchylka :	0.1020426907
Medianový rozpty :	0.01041271072
10% Průměr :	3.895227273
10% IS spodní :	3.835978432
10% IS horní :	3.954476113
10% Směr. odchylka :	0.149154143
10% Rozptyl :	0.02224695837
20% Průměr :	3.897
20% IS spodní :	3.835649773
20% IS horní :	3.958350227
20% Směr. odchylka :	0.1243187477
20% Rozptyl :	0.01545515102
40% Průměr :	3.901666667
40% IS spodní :	3.838030654
40% IS horní :	3.96530268
40% Směr. odchylka :	0.08167506551
40% Rozptyl :	0.006670816327

Robustní parametry:

Medián Odhad mediánu, tedy 50% kvantilu. Tento odhad střední hodnoty je spolehlivější než aritmetický průměr v případě porušení normality dat nebo přítomnosti vybočujících bodů.
IS spodní Spodní mez intervalu spolehlivosti mediánu na zadané hladině významnosti.
IS horní Horní mez intervalu spolehlivosti mediánu na zadané hladině významnosti.
Mediánová směr.odchylka Odhad směrodatné odchylky na základě mediánu.
Medianový rozptyl Odhad rozptylu na základě mediánu.

10% ořezaný průměr Aritmetický průměr pro symetrickém ořezání 10% dat, tedy 5% nejmenších a 5% největších hodnot. Tento robustní odhad střední hodnoty se doporučuje v případě podezření na vybočující body.
IS spodní Spodní mez intervalu spolehlivosti ořezaného průměru na zadané hladině významnosti.
IS horní Horní mez intervalu spolehlivosti ořezaného průměru na zadané hladině významnosti.
Rozptyl Odhad rozptylu na zálkladě mediánu.
Směr. odchylka Odhad směrodatné odchylky na základě mediánu.

"40% ořezaný průměr Aritmetický průměr pro symetrickém ořezání 40% dat, tedy 20% nejmenších a 20% největších hodnot. Tento robustní odhad střední hodnoty se doporučuje v případě podezření na velký počet vybočujících bodů.
IS spodní Spodní mez intervalu spolehlivosti ořezaného průměru na zadané hladině významnosti.
IS horní Horní mez intervalu spolehlivosti ořezaného průměru na zadané hladině významnosti.
Rozptyl Odhad rozptylu na zálkladě mediánu.
Směr. odchylka Odhad směrodatné odchylky na základě mediánu.

Znaménkový test :

Závěr :	Data jsou nezávislá

Znaménkový test Neparametrický test nezávislosti dat, závislost je indikována, obsahuji-li data shluky po sobě jdoucích sekvencí se shodným znaménkem odchylky od průměru.
Závěr Slovní závěr testu zavislosti dat.

Test normality :

Název sloupce :	B201
Průměr :	3.894
Rozptyl :	0.0394
Šikmost	-0.119116936
Špičatost :	2.40159769
Normalita :	Přijata
Vypočtený :	0.2378781079
Teoretický :	5.991464547
Pravděpodobnost :	0.8878619108

Test normality Kombinovaný test normality založený na shodě šikmosti a špičatosti s normálním rozdělením.

Normalita Slovní závěr testu na zadané hladině významnosti.

Vypočtený Vypočtená testovací statistika.

Teoretický Příslušný kvantil t-rozdělení.

Pravděpodobnost Pravděpodobnost odpovídající vypočtené statistice.

Vybočující body :

Název sloupce :	B201
Homogenita :	Přijata
Počet vybočujících bodů :	0
Spodní mez :	3.04482
Horní mez :	4.70518

Vybočující body Robustní test na přítomnost vybočujících měření založený na kvantilovém odhadu vnitřních mezí dat.

Homogenita Slovní závěr testu, nejsou-li v datech vybočující měření, je předpoklad homogenity přijat.

Počet vybočujících bodů Počet případných měření přesahujících přípustné meze, které je možno považovat za vybočující.

Dolní hranice Dolní hranice, pod níž je možno data považovat za vybočující.

Horní hranice Horní hranice, nad níž je možno data považovat za vybočující.

Autokorelace :

Řád autokorelace :	4
Název sloupce :	B201
Počet :	0.0521889568
Řád autokorelace 1	
Korelační koeficient :	0.01257847301
Pravděpodobnost :	0.4658208006
Závěr :	Nevýznamný
Řád autokorelace 2	
Korelační koeficient :	-0.1020884699
Pravděpodobnost :	0.2449546948
Závěr :	Nevýznamný
Řád autokorelace 3	
Korelační koeficient :	-0.02901067026
Pravděpodobnost :	0.4232554223
Závěr :	Nevýznamný
Řád autokorelace 4	
Korelační koeficient :	0.0521889568
Pravděpodobnost :	0.3652515549
Závěr :	Nevýznamný

Autokorelace Odhady autokorelačních koeficientů a jejich významnost na zadané hladině významnosti.

Řád autokorelace Řád autokorelace.

Koeficient Hodnota autokorelačního koeficientu, formálně odpovídá párovému korelačnímu koeficientu a má stejné vlastnosti.
Pravděpodobnost Pravděpodobnost nevýznamnosti autokorelační-ho koeficientu; je-li menší než zvolená hladina významnosti, je autokorelace významná.
R0Krit Kritická hodnota autokorelačního koeficientu, nad níž se korelace považuje za významnou.
Vysledek Slovní vyjádření významnosti autokorelace.
Vyhlazené hodnoty Hodnoty klouzavých průměrů a mediánů. Při vhodné volbě řádu trendu lze získat po odečtení těchto hodnot od vstupních dat detrendovaná data, která lze použít pro konstrukci regulačních diagramů v případě, že je v datech příliš silný trend.
Rezidua Odchylky (rozdíly) naměřených a vyhlazených hodnot, (naměřená - vyhlazená).

Test významnosti trendu :

Název sloupce :	B201
Směrnice :	0.0004225690276
Významnost :	Nevýznamný
Pravděpodobnost :	0.5847032011

Test významnosti trendu Test významnosti lineárního trendu v datech.

Vypočtený Testovací statistika pro směrnici přímky vypočtená z dat.
Teoretický Příslušný kvantil t-rozdělení.
Trend Hodnota směrnice přímky proložené daty
Pravděpodobnost Pravěpodobnost toho, že je lineární trend nevýznamný; vyjde-li menší než zadaná hladina významnosti (tedy

Závěr EDA:

1. Rozdělení je symetrické (mezi rovnoměrným a Gaussovým).

2. Rozdělení obsahuje nahoře 2 a dole cca 5 odlehlých bodů – outlierů.

3. Není asi nutné užít transformaci dat.

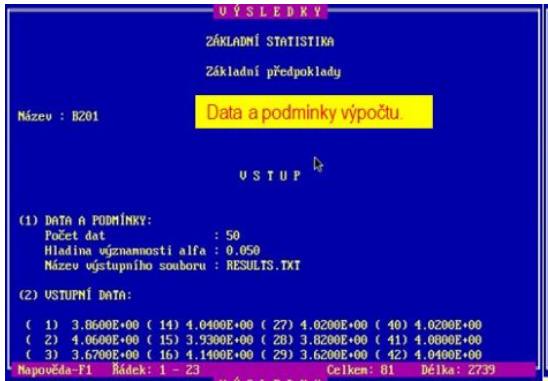
2. kapitola
EDA: Ověření předpokladů o výběru

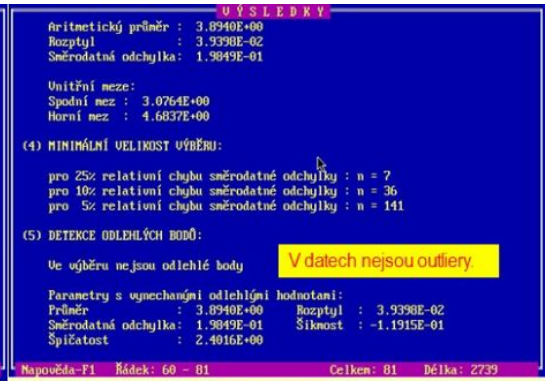
Řešení úloh z Kompendia:

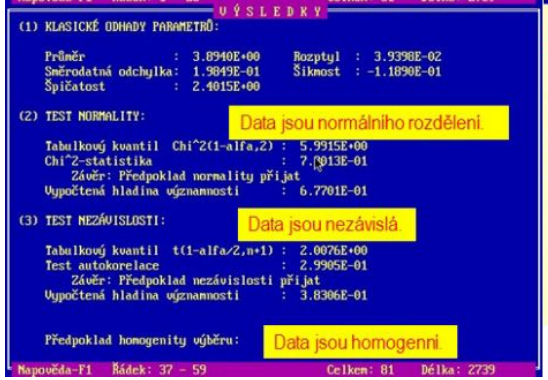
B201

2. Ověření předpokladů výběru dat:

Diagnosticky, testy: *ověření normality*
ověření nezávislosti
ověření homogenity
určení minimální četnosti



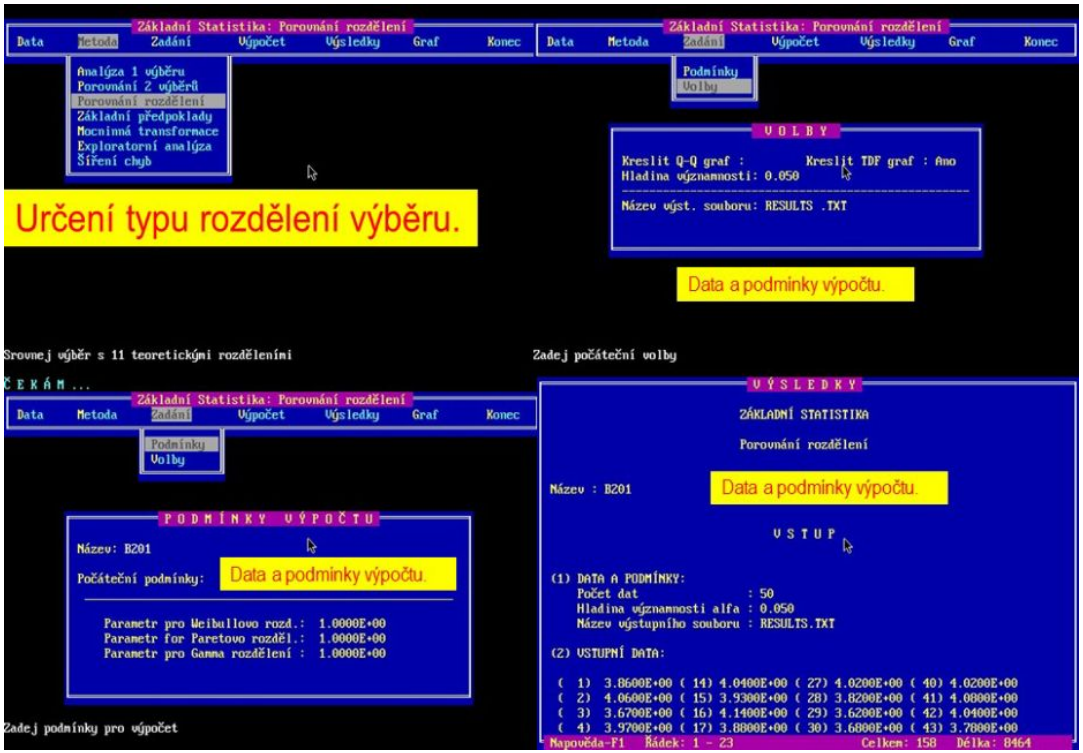


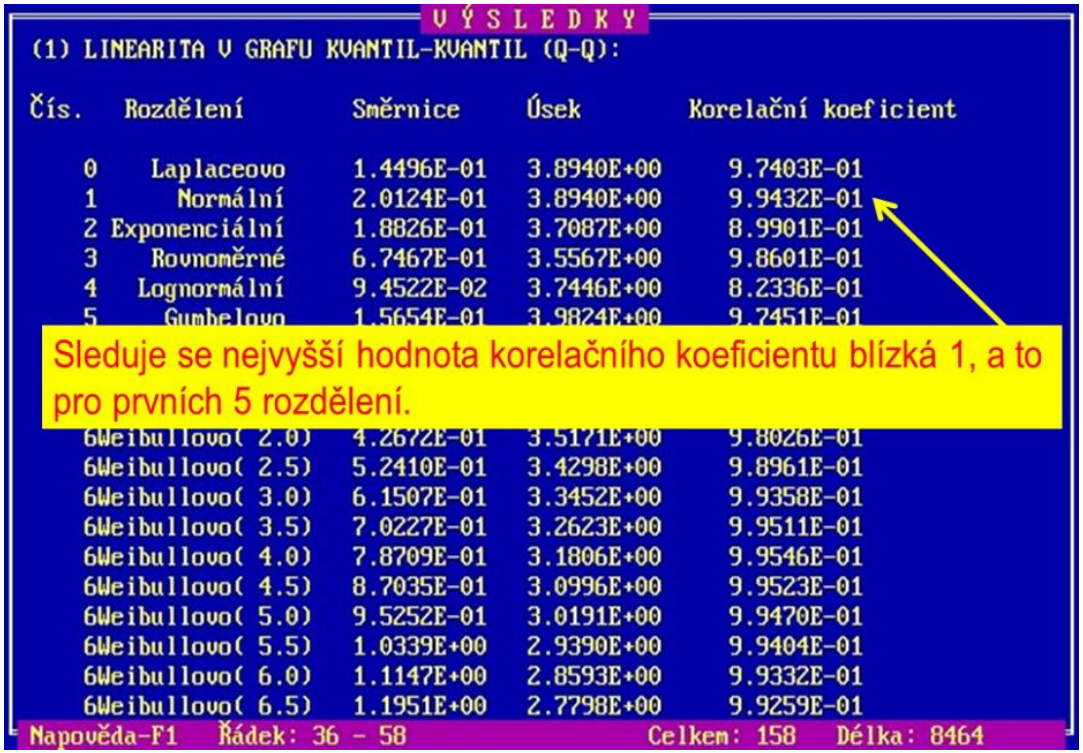


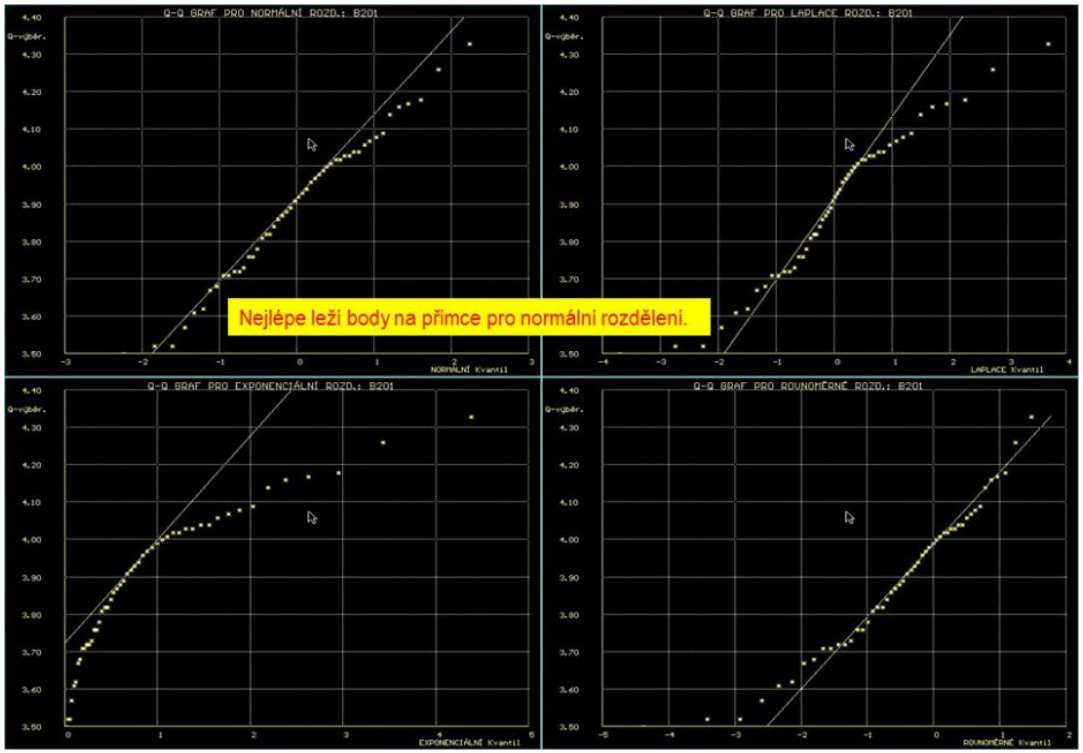
Závěr Základních předpokladů:

Data vykazují Gaussovo normální rozdělení a není proto třeba použít jakoukoliv transformaci dat.

Test prokázal, že data jsou nezávislá a bez vybočujících hodnot.







2. kapitola

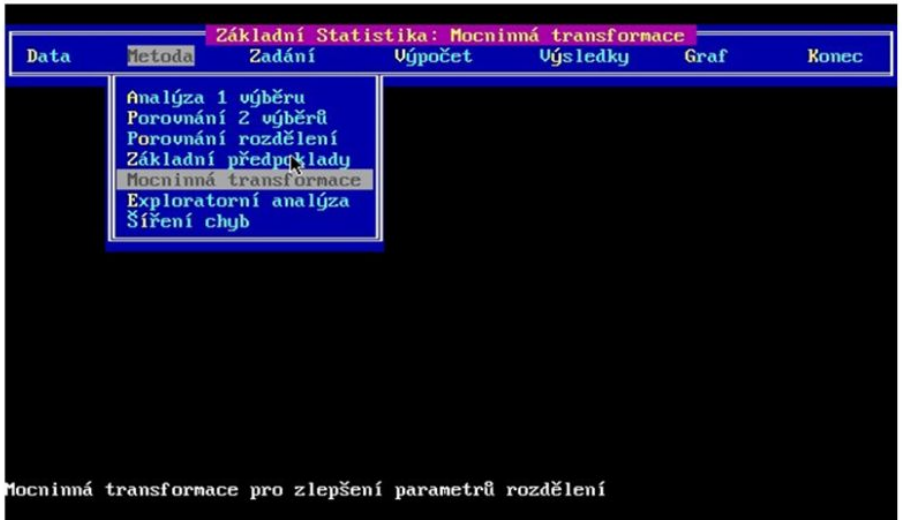
EDA: Transformace dat

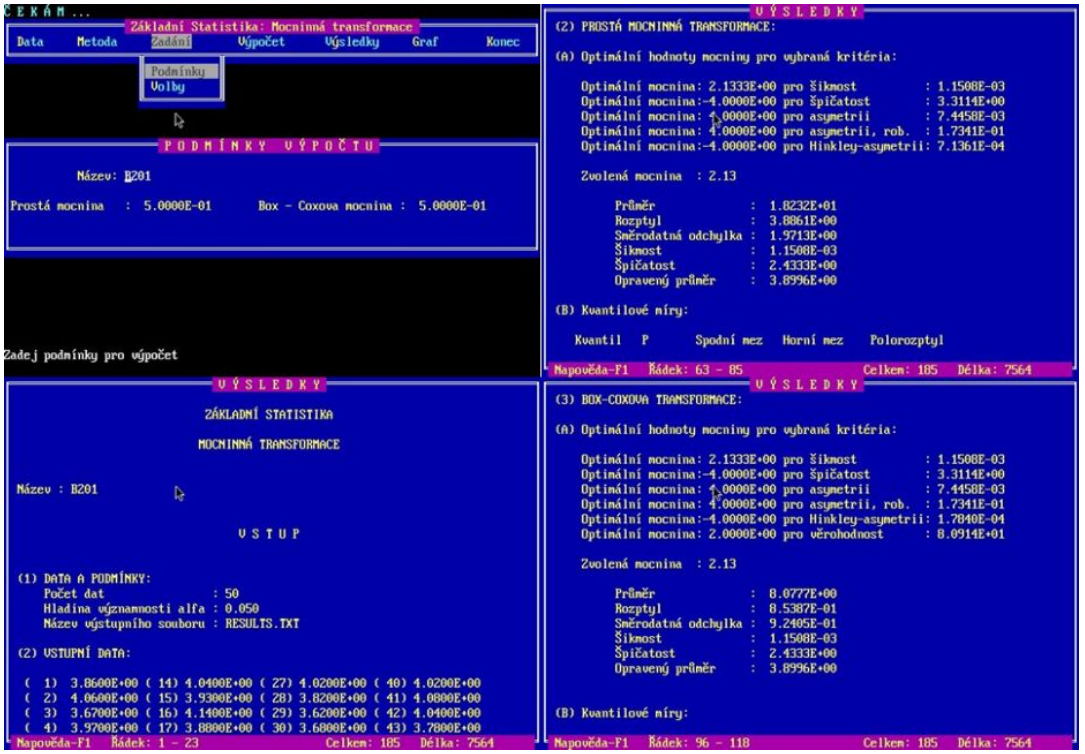
Řešení úloh z Kompendia:

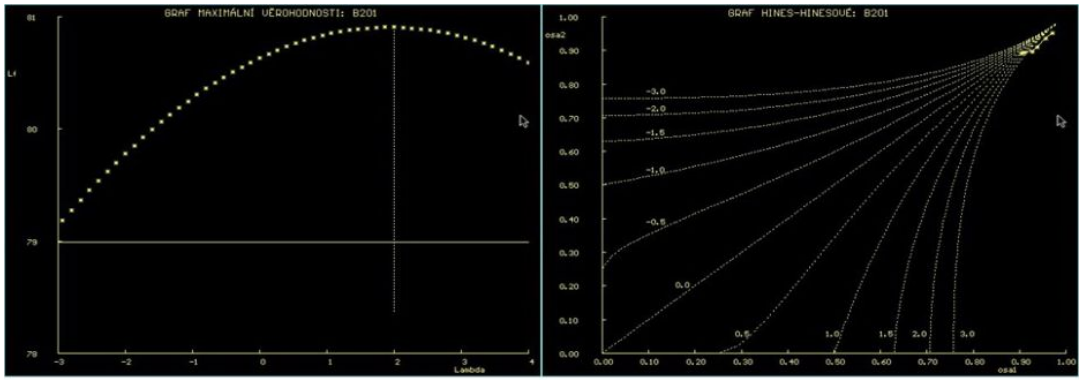
B201

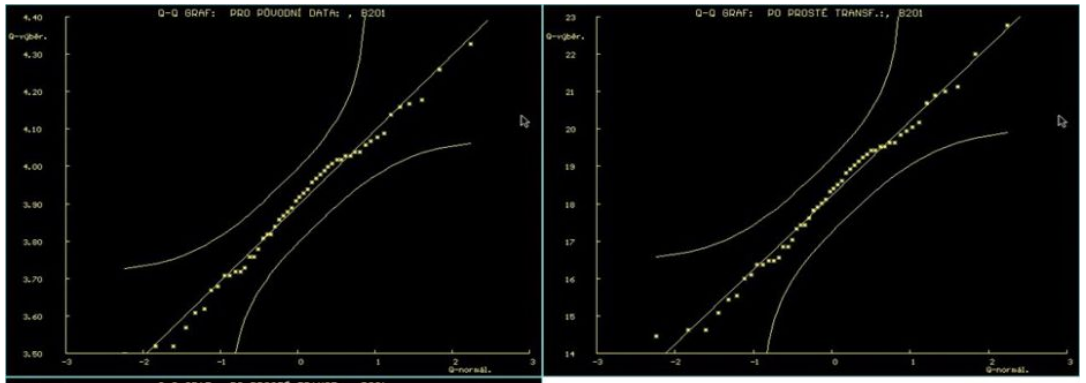
3. Transformace dat:

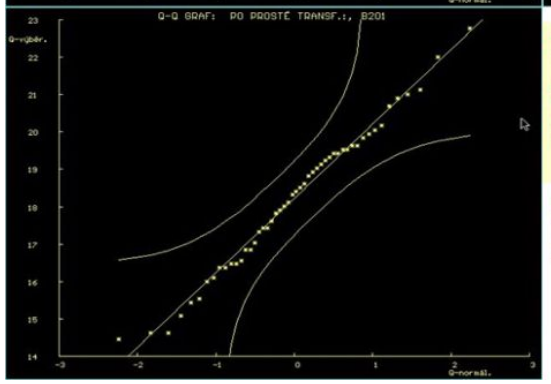
Analýza dat: *originální data*
data po mocninné transformaci
data po Box-Coxově transformaci











Rozhodčí kritérium použití transformace dat:

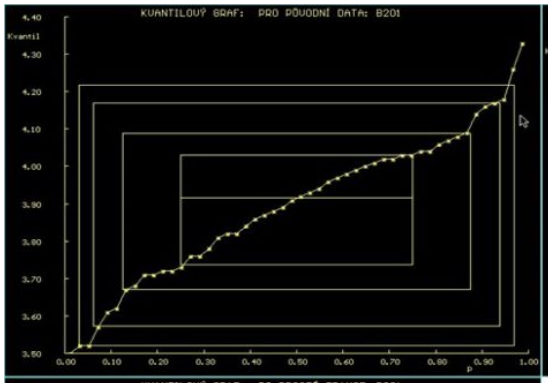
Nachází-li se pod segmentem na x-ové ose číslo 1 , pak je transformace zbytečná a lze využít pro odhad střední hodnoty aritmetický průměr.

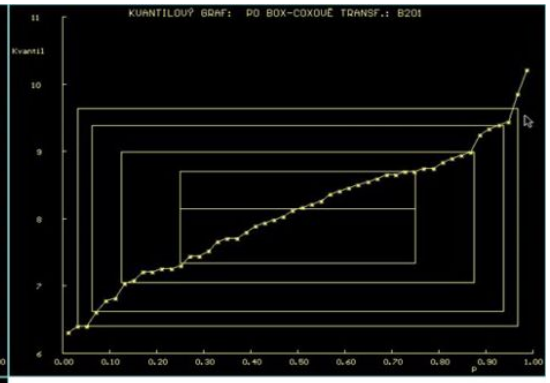
Určení nejlepšího odhadu exponentu lambda grafickou metodou dle Hinese a Hinesové:

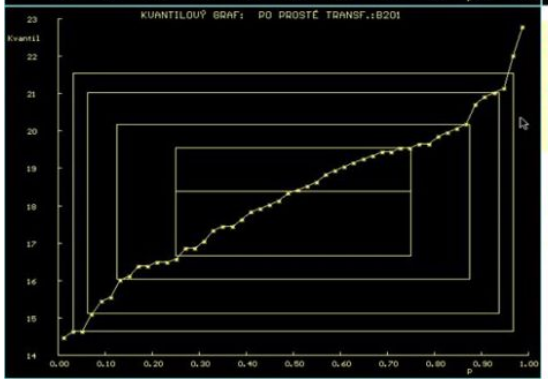
V nomogramu Hines-Hines se odhadne poloha křivky s body, která se přibližuje některému průvodiči v nomogramu.

Ukázka postupného zlepšení symetrie rozdělení po transformaci dat.

Symetrie je indikovaná v Q-Q kvantil-kvantilovém grafu.

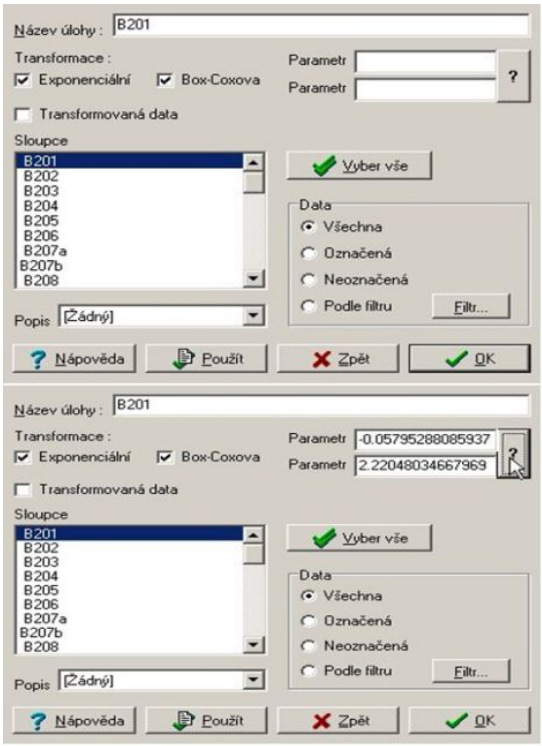


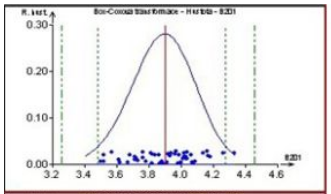




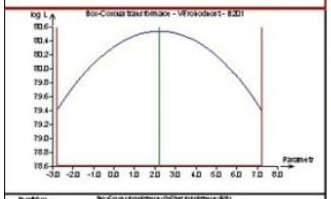
Ukázka postupného zlepšení symetrie rozdělení po transformaci dat.

Symetrie je indikovaná v grafu rozptýlení s kvantily.

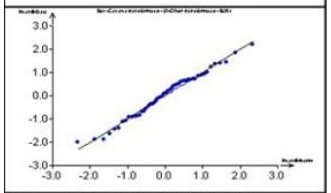




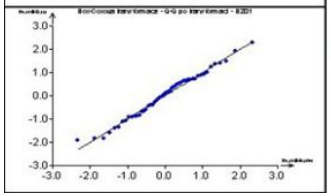
Graf hustoty představuje tvar rozdělení, který nejlépe vystihuje data prostředky Box-Coxovy transformace. Svislé čáry představují kvantily (hodnoty) odpovídající (od středu) mediánu (50% kvantil), kvartilu (25% kvantily ohraničující 50% dat), ±2s (zhruba 2.5% kvantily ohraničující interval 95% dat), 0.5% kvantily ohraničující 99% dat a ±3s (ohraničující 99.73% dat).



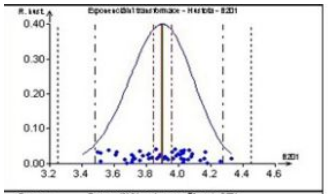
Graf logaritmu závislosti věrohodnostní funkce (osa y) na exponentu lambda. Maximu odpovídá optimální hodnota lambda. Vodorovná přímka odpovídá spodní mezi 95% intervalu spolehlivosti maxima věrohodnosti a svislé přímky odpovídají intervalu spolehlivosti odhadu lambda. Obsahuje-li tento interval 1, není nutné transformovat a je možné použít odhady v Základní statistice, případně transformaci s lambda = 1.



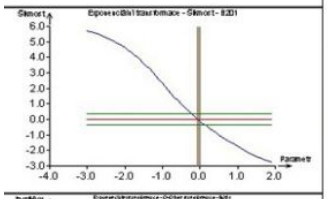
QQ-graf původních dat, shodný s QQ-grafem v Základní statistice. Metoda transformace bývá zpravidla užitečný jen pro systematicky prohnutý tvar bodů v QQ-grafu, viz vlevo. Proti statistikám má QQ-graf výhodu v možnosti vizuálně posoudit, zda je nelinearita (tedy odchylka od normality) způsobena jen několika body, nebo všemi daty.



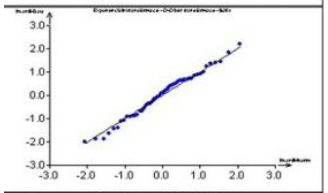
QQ-graf dat po transformaci: Je-li tvar bodů blíže přímce než na předešlém grafu, je transformace úspěšná. Ke kvantitavnímu posouzení je však třeba použít statistik uvedených v protokolu.



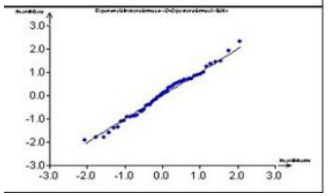
Graf hustoty představuje tvar rozdělení, který nejlépe vystihuje data prostředky Box-Coxovy transformace. Svislé čáry představují kvantily (hodnoty) odpovídající (od středu) mediánu (50% kvantil), kvartilu (25% kvantily ohraničující 50% dat), ±2s (zhruba 2.5% kvantily ohraničující interval 95% dat), 0.5% kvantily ohraničující 99% dat a ±3s (ohraničující 99.73% dat).



Závislost šikmosti transformovaných dat na parametru transformace: Nulová šikmost odpovídá optimálnímu parametru. Význam tohoto grafu je podobný jako u předchozího grafu věrohodnosti, slouží k nalezení parametru transformace a určení statistické významnosti transformace. Leží-li průsečík svislé zelené přímky s křivkou mimo interval spolehlivosti šikmosti (vodorovné zelené přímky), je transformace opodstatněná.



QQ-graf původních dat, shodný s QQ-grafem v Základní statistice. Metoda transformace bývá zpravidla užitečný jen pro systematicky prohnutý tvar bodů v QQ-grafu, viz vlevo. Proti statistikám má QQ-graf výhodu v možnosti vizuálně posoudit, zda je nelinearita (tedy odchylka od normality) způsobena jen několika body, nebo všemi daty.



QQ-graf dat po transformaci: Je-li tvar bodů blíže přímce než na předešlém grafu, je transformace úspěšná. Ke kvantitavnímu posouzení je však třeba použít statistik uvedených v protokolu.

Box-Coxova transformace dat :

Název úlohy :	B201
Data:	Všechna
Vybrané sloupce :	
B201	

Optimální parametr :	2.220480347
Dolní mez parametru :	-2.779519653
Horní mez parametru :	7.220480291
Věrohodnost bez transformace :	80.47189562
Věrohodnost s transformací :	80.53909592
Oprávněnost transformace :	Ne
Pravděpodobnost :	28.6029656815284 %
Zvolený parametr :	2.220480347
Věrohodnost :	80.53909592
Opravený průměr :	3.898743595
LCL :	3.254893143
UCL :	4.453982689
LWL :	3.695897375
UWL :	4.091950865

Optimální parametr Nejlepší doporučená hodnota parametru r, při niž je dosaženo nejlepší shody s normálním rozdělením na základě maximální věrohodnosti.

Dolní a horní mez parametru Interval spolehlivosti optimální hodnoty r. Hodnoty uvnitř tohoto intervalu poskytnou podobný efekt jako hodnota optimální. Interval se obvykle zužuje s rostoucím počtem dat. Je-li uvnitř tohoto intervalu jednička, není účelné data transformovat buď z toho důvodu, že data již normální jsou, nebo je dat příliš málo, interval spolelivosti r je příliš široký a není možné najít jednoznačnou transformaci. Je-li uvnitř intervalu nula, lze rozdělení dat považovat za lognormální.

Věrohodnost bez transformace Hodnota logaritmu věrohodnosti normality netransformovaných dat vzhledem k normálnímu rozdělení. Protože se jedná o logaritmy, odpovídá rozdílu o 1 řádovému rozdílu.

Věrohodnost transformací Hodnota logaritmu věrohodnosti po transformaci s optimálním parametrem, tedy maximální dosažitelná shoda s normálním rozdělením. Protože se jedná o logaritmy, odpovídá rozdílu o 1 řádovému rozdílu.

Oprávněnost transformace Slovní vyjádření opodstatnění transformace. NE znamená, že transformace neposkytuje významný přínos. ANO znamená doporučení transformace. Transformace se doporučuje, je-li účinnost transformace vyšší než 95%.

Účinnost transformace Statistická významnost transformace. Je v podstatě kvantitativním vyjádřením oprávněnosti transformace (předchozí položka). Je-li účinnost větší než 95%, považuje se transformace za oprávněnou a doporučenou, jinak se transformace nedoporučuje. Hranice 95% však není striktní. Je-li účinnost blízká 95%, lze rozhodnout i opačně.

Zvolený parametr Hodnota uživatelem zadaného parametru v dialogovém panelu. Tato hodnota může být různá od doporučené optimální hodnoty.

Věrohodnost Logaritmus věrohodnosti odpovídající zvolenému parametru.

Opravený průměr Aritmetický průměr vypočítaný metodou Box-Coxovy transformace. V případě asymetrických dat odpovídá střední hodnotě lépe než prostý průměr počítaný v Základní statistice.

LCL Doporučená hodnota spodní kontrolní meze pro případ konstrukce Shewhartova regulačního diagramu typu X-průměr. Počet sloupců je chápán jako velikost podskupiny.

UCL Doporučená hodnota spodní kontrolní meze pro případ konstrukce Shewhartova regulačního diagramu typu X-průměr. Počet sloupců je chápán jako velikost podskupiny.

LWL Doporučená hodnota spodní varovné meze.

UWL Doporučená hodnota horní varovné meze.

Exponenciální transformace dat :

Název úlohy :	B201
Data:	Všechna
Vybrané sloupce :	
B201	

Optimální parametr :	-0.05795288086
Zvolený parametr :	-0.05795288086
Oprávněnost transformace :	Ne
Opravený průměr :	3.899547054
Interval spolehlivosti :	
Spodní :	3.842812816
Horní :	3.95535866
LCL :	3.250862643
UCL :	4.447294615
LWL :	3.479724116
UWL :	4.274265575

Optimální parametr Nejlepší doporučená hodnota parametru r, při niž je dosaženo nejlepší shody s normálním rozdělením na základě maximální věrohodnosti.

Dolní a horní mez parametru Interval spolehlivosti optimální hodnoty r. Hodnoty uvnitř tohoto intervalu poskytnou podobný efekt jako hodnota optimální. Interval se obvykle zužuje s rostoucím počtem dat. Je-li uvnitř tohoto intervalu jednička, není účelné data transformovat buď z toho důvodu, že data již normální jsou, nebo je dat příliš málo, interval spolelivosti r je příliš široký a není možné najít jednoznačnou transformaci. Je-li uvnitř intervalu nula, lze rozdělení dat považovat za lognormální.

Věrohodnost bez transformace Hodnota logaritmu věrohodnosti normality netransformovaných dat vzhledem k normálnímu rozdělení. Protože se jedná o logaritmy, odpovídá rozdílu o 1 řádovému rozdílu.

Věrohodnost transformací Hodnota logaritmu věrohodnosti po transformaci s optimálním parametrem, tedy maximální dosažitelná shoda s normálním rozdělením. Protože se jedná o logaritmy, odpovídá rozdílu o 1 řádovému rozdílu.

Oprávněnost transformace Slovní vyjádření opodstatnění transformace. NE znamená, že transformace neposkytuje významný přínos. ANO znamená doporučení transformace. Transformace se doporučuje, je-li účinnost transformace vyšší než 95%.

Účinnost transformace Statistická významnost transformace. Je v podstatě kvantitativním vyjádřením oprávněnosti transformace (předchozí položka). Je-li účinnost větší než 95%, považuje se transformace za oprávněnou a doporučenou, jinak se transformace nedoporučuje. Hranice 95% však není striktní. Je-li účinnost blízká 95%, lze rozhodnout i opačně.

Zvolený parametr Hodnota uživatelem zadaného parametru v dialogovém panelu. Tato hodnota může být různá od doporučené optimální hodnoty.

Věrohodnost Logaritmus věrohodnosti odpovídající zvolenému parametru.

Opravený průměr Aritmetický průměr vypočítaný metodou Box-Coxovy transformace. V případě asymetrických dat odpovídá střední hodnotě lépe než prostý průměr počítaný v Základní statistice.

LCL Doporučená hodnota spodní kontrolní meze pro případ konstrukce Shewhartova regulačního diagramu typu X-průměr. Počet sloupců je chápán jako velikost podskupiny.

UCL Doporučená hodnota spodní kontrolní meze pro případ konstrukce Shewhartova regulačního diagramu typu X-průměr. Počet sloupců je chápán jako velikost podskupiny.

LWL Doporučená hodnota spodní varovné meze.

UWL Doporučená hodnota horní varovné meze.

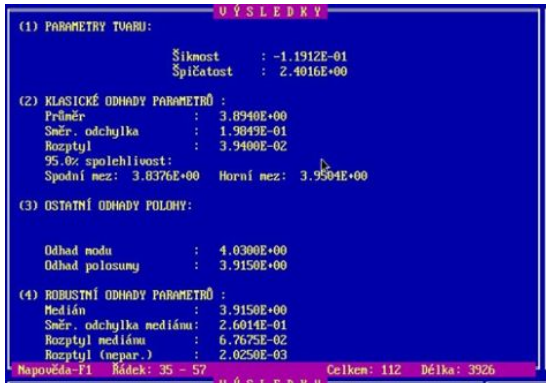
4. Parametry polohy, rozptýlení a tvaru:

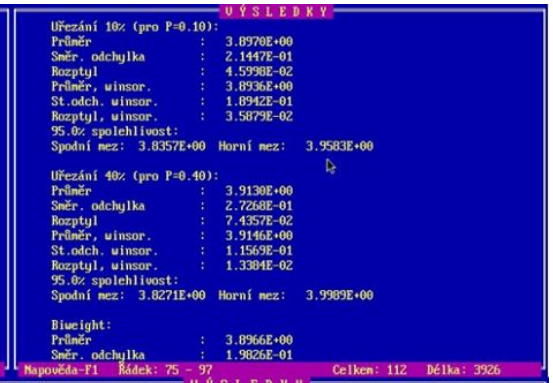
Analýza 1 výběru:

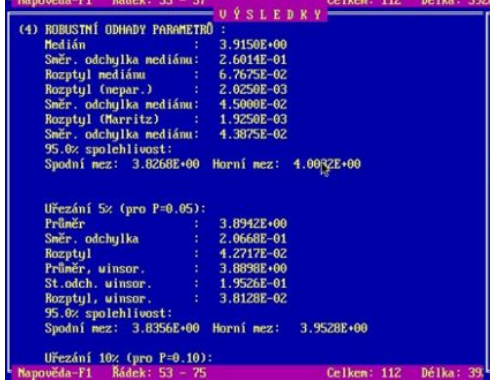
klasické odhady
- *průměr*
- *rozptyl*

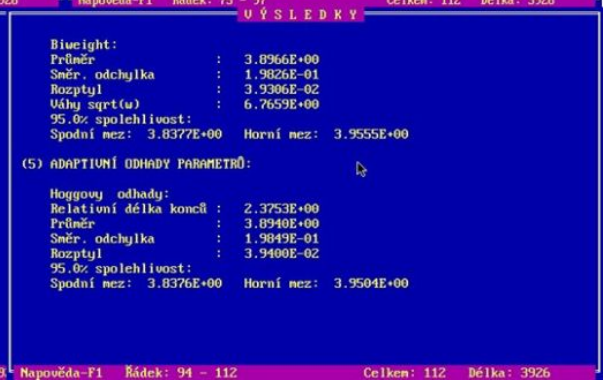
robustní odhady
- *medián*
- *uřezané průměry*
- *winsorizovaný rozptyl*
- *interkvantilové rozpětí*

adaptivní odhady









Závěr:

Protože EDA a základní předpoklady prokázaly symetrické normální rozdělení hustoty pravděpodobnosti, lze za nejlepší odhad polohy použít bodový a intervalový odhad aritmetického průměru. Retransformovaný průměr vede totiž k numericky stejné hodnotě.