

## 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*

# 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

## 2. kapitola EDA: Exploratorní analýza dat

Řešení úloh z Kompendia:

# B204

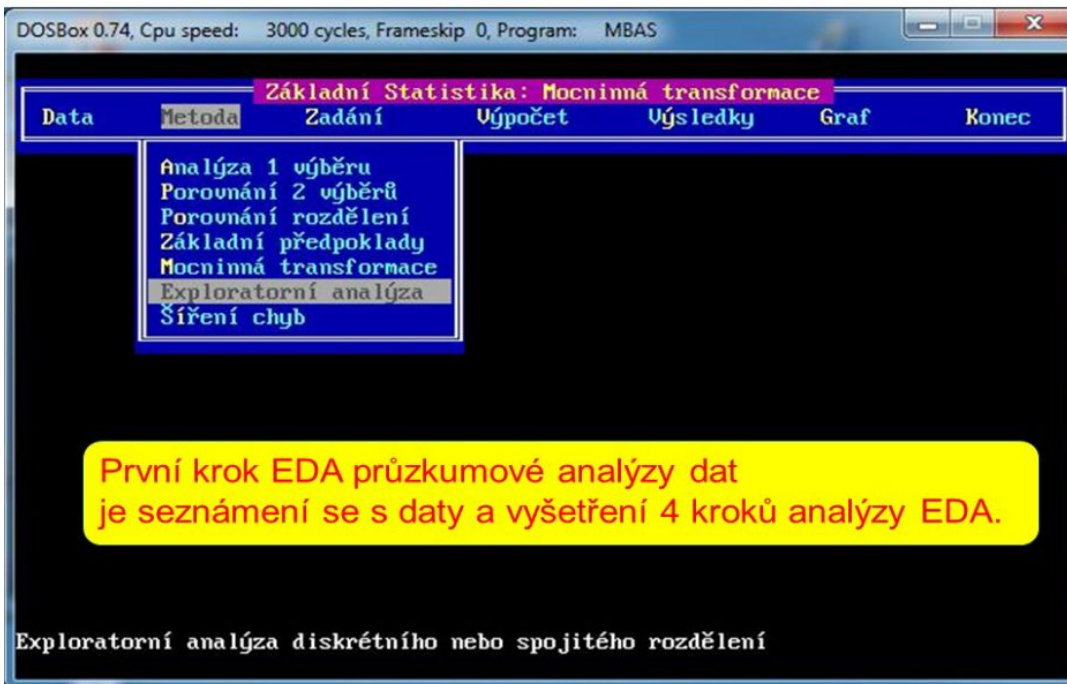
## Úloha B2.04 Typ rozdělení a odlehle body obsahu ergosterinu v kalciferolu

**Zadání:** Při výrobě kalciferolu se provádí kontrola meziprojektu 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 kvantily 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 [%]


**Volba metody:** Exploratorní analýza dat



První krok EDA průzkumové analýzy dat je seznámení se s daty a vyšetření 4 kroků analýzy EDA.

Exploratorní analýza diskrétního nebo spojitého rozdělení

**Zadání dat:** formou kódovaného čísla úlohy z Kompendia



Po F3 napíšeme číslo úlohy z Kompendia Formátem B204 nebo B204.txt nebo b204.

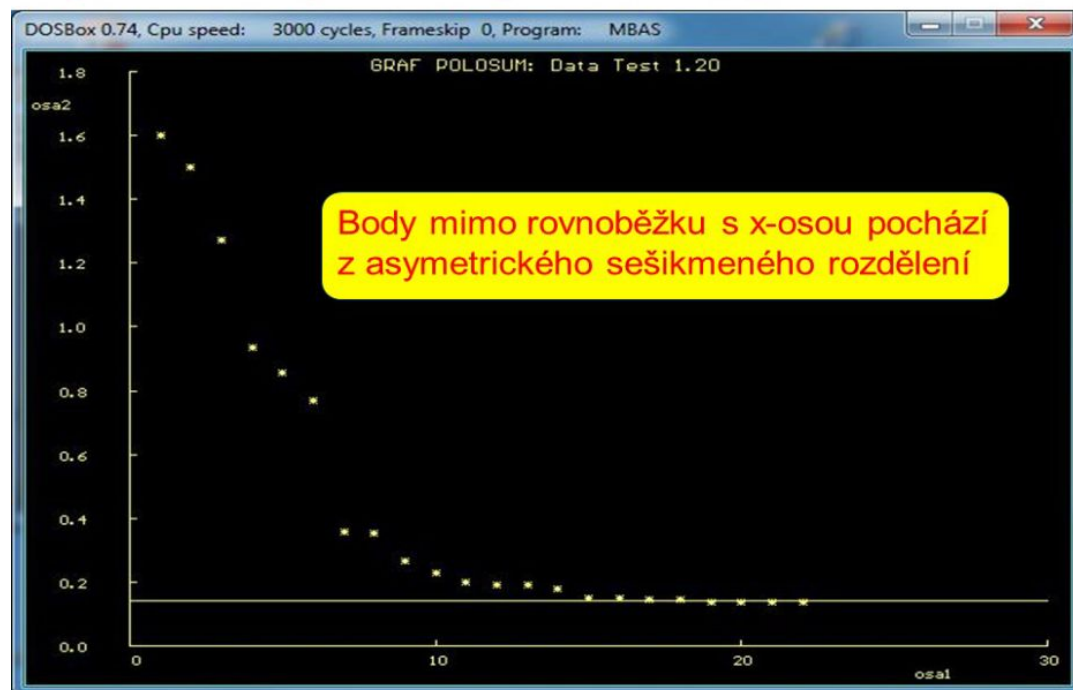
**Diagnostikování:** Kvantilový graf ukazuje na dlouhý ocas a sešikmené rozdělení



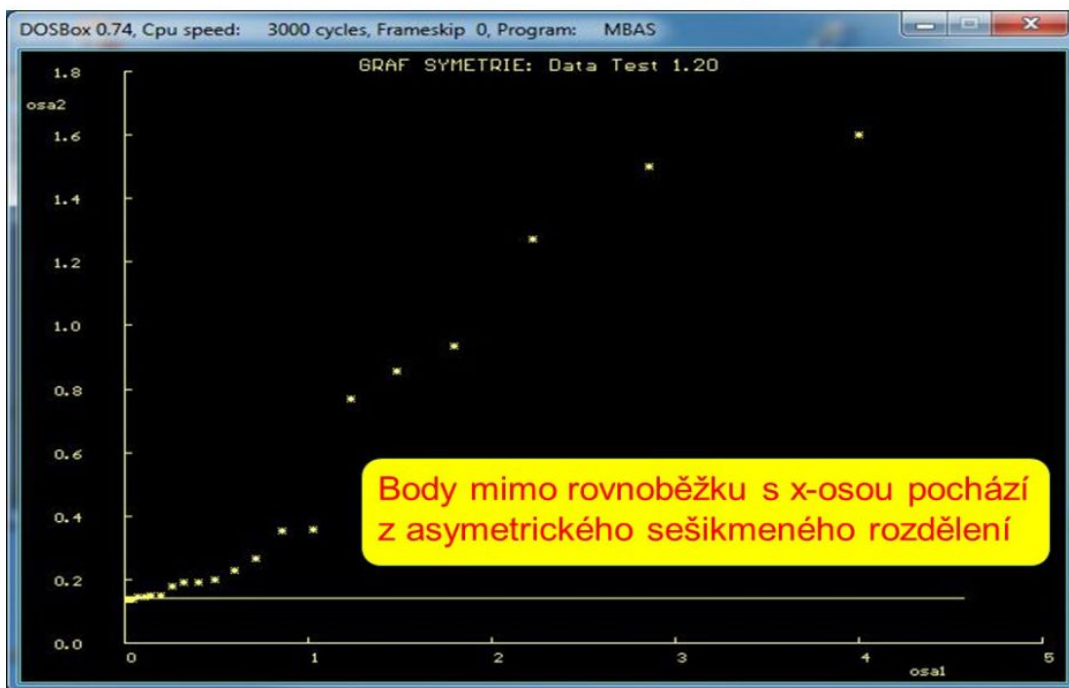
**Diagnostikování:** Bodové grafy ukazují řadu odlehlých bodů nahore



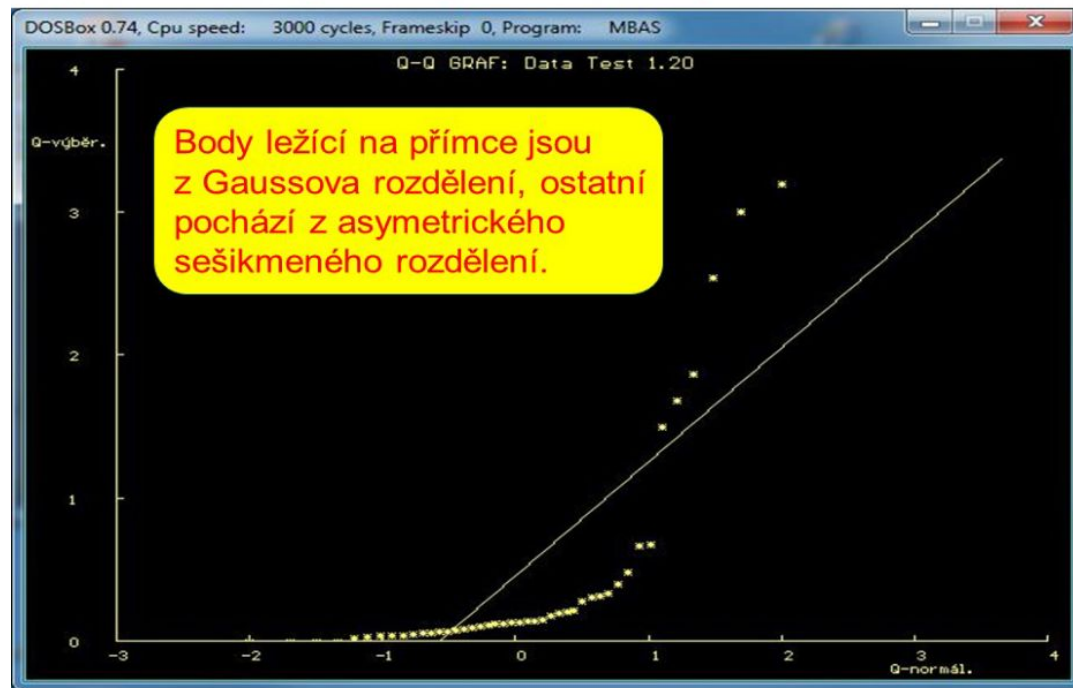
**Diagnostikování:** Graf polosum odhaluje silně sešikmené, asymetrické rozdělení



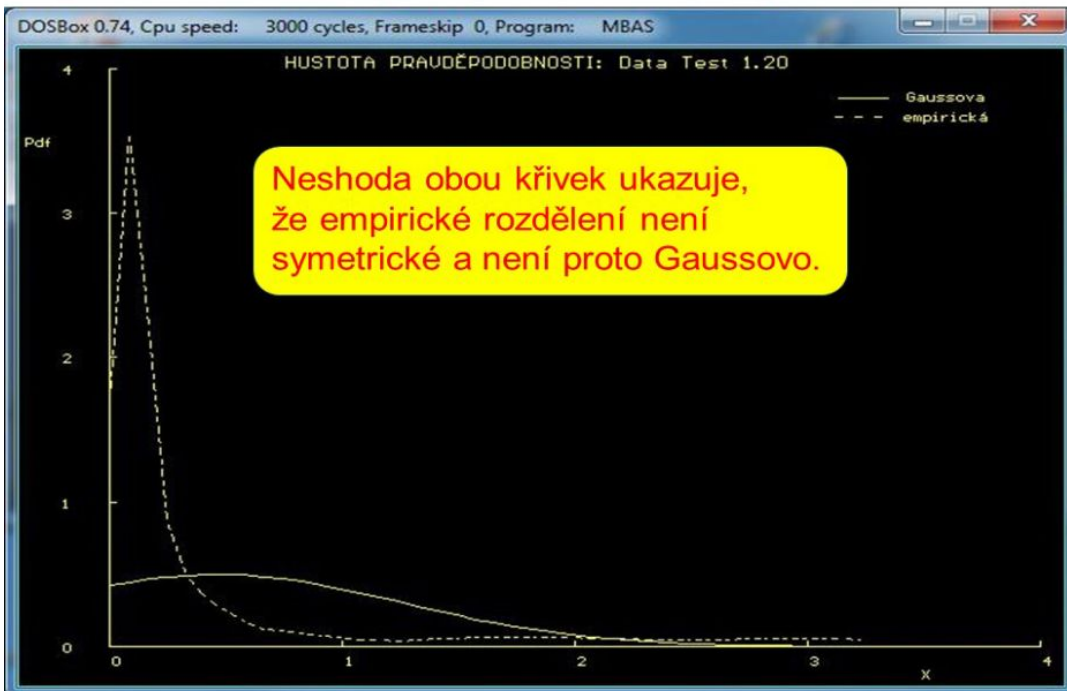
**Diagnostikování:** Graf symetrie odhaluje silně sešikmené, asymetrické rozdělení



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



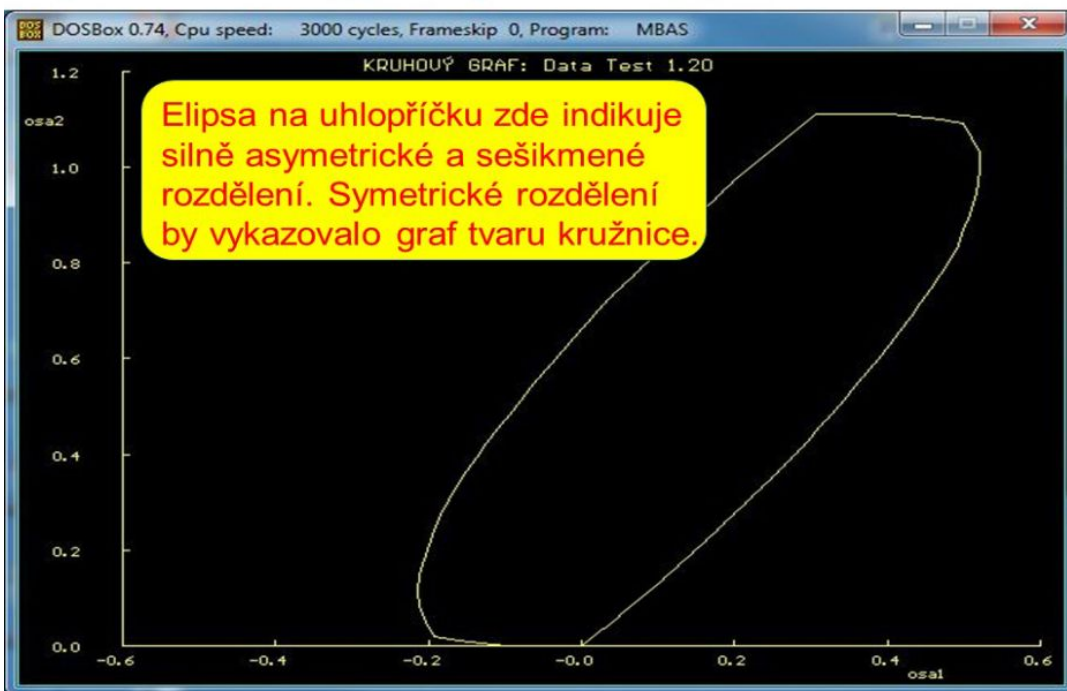
**Diagnostikování:** Graf hustoty pravděpodobnosti ukazuje na špičaté asymetrické r.



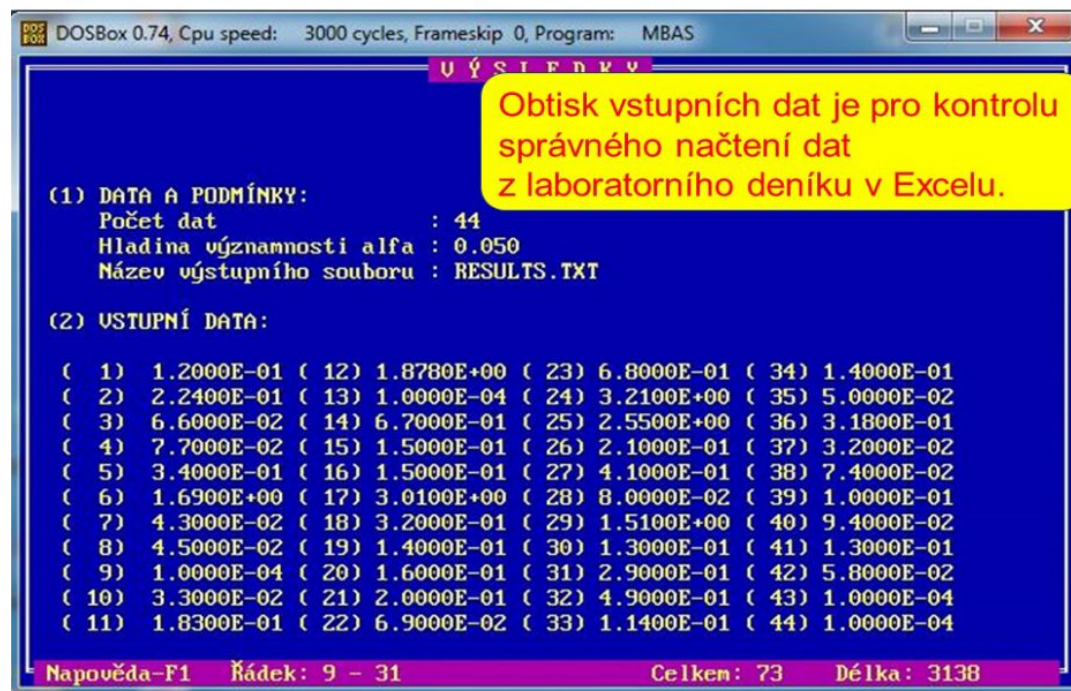
**Diagnostikování:** Graf rozptýlení s kvantily ukazuje na sešikmené, asymetrické r.



**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

```
DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS
U Ý S L E D K Y
(1) KLASICKÉ ODHADY PARAMETRŮ
Medián      : 1.4000E-01      Průměr      : 4.5996E-01
Rozptyl     : 6.3581E-01      III.cent. moment : 1.1816E+00
IV.cent. moment : 2.9474E+00      Šikmost     : 2.3575E+00
Špičatost  : 7.4605E+00      Směrodatná odchylka: 7.9738E-01

(2) KVANTILY A PÍSMENOVÉ HODNOTY
Kvantilové míry:

Procento   Kvantil                               Procento   Kvantil
5          1.0000E-04                               10         3.2300E-02
15         4.3900E-02                               20         5.4800E-02
25         6.8250E-02                               30         7.6700E-02
35         9.4300E-02                               40         1.1520E-01
45         1.3000E-01                               50         1.4000E-01
55         1.5000E-01                               60         1.7840E-01
65         2.0950E-01                               70         2.9280E-01
75         3.2500E-01                               80         4.4200E-01
85         6.7550E-01                               90         1.6360E+00
95         2.4492E+00

Napověda-F1  Řádek: 34 - 56                               Celkem: 73  Délka: 3138
```

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

**Output:** Písmenové hodnoty a kvantilové míry tří parametrů

```
DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS
U Ý S L E D K Y
Písmenové hodnoty:
Kvantil Písmeno  Pravděpodobnost  Spodní mez  Horní mez
Sedecil D      0.0625           1.0000E-04  2.0880E+00
Oktil   E      0.1250           3.6750E-02  1.1987E+00
Kvartil F      0.2500           6.8250E-02  3.2500E-01
Medián  M      0.5000           1.4000E-01  1.4000E-01

(3) KVANTILOVÉ MÍRY:
Kvantil   :      F (0.25)           E (0.125)           D(0.0625)
Rozsah    :      2.5675E-01           1.1620E+00           2.0879E+00
Polosuma  :      1.9662E-01           6.1775E-01           1.0440E+00
Délka konců: 0.0000E+00           1.5098E+00           2.0958E+00
Šikmost   :      -1.9656E-01           -3.1430E-01           -2.5974E-01
PseudoSigma: 1.9047E-01           5.0522E-01           6.8232E-01

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

Napověda-F1  Řádek: 57 - 73                               Celkem: 73  Délka: 3138
```

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.

## 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  $\rho_a$ :**

**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  $\alpha$  zamítnout.

## Ověření nezávislosti prvků výběru

Závislost měření je způsobena:

- nestabilitou měřicího zařízení,
- nekonstantností podmínek měření,
- zanedbáním faktorů: objem vzorků, teplota, nečistota, ...
- nesprávným, nenáhodným výběrem vzorků k měření,
- 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  $C_1 > \chi_{1-\alpha}^2(2)$ , je nutno hypotézu o normalitě rozdělení výběru zamítnout.

# Ověření homogenity výběru

Modifikace vnitřních hradeb  $B_D^*$  a  $B_H^*$

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

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

**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.

## 2. kapitola EDA: Ověření předpokladů o výběru

Řešení úloh z Kompendia:

# B204



```
DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS
Základní Statistika: Exploratorní analýza spojitá
Data Metoda Zadání Úpočet Výsledky Graf Konec
Analýza 1 výběru
Porovnání 2 výběrů
Porovnání rozdělení
Základní předpoklady
Mocninná transformace
Exploratorní analýza
Šíření chyb
Testování splnění základních předpokladů o datech
```

```
DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS
Klíčově důležitý test normality ukazuje, že jediným řešením je transformace dat.
(1) KLASICKÉ ODHADY PARAMETRŮ:
Průměr : 4.5996E-01 Rozptyl : 6.3580E-01
Směrodatná odchylka : 7.9737E-01 Šikmost : 2.3576E+00
Špičatost : 7.4607E+00
(2) TEST NORMALITY:
Tabulkový kvantil Chi^2(1-alfa,2) : 5.9915E+00
Chi^2-statistika : 1.0078E+02
Závěr: Předpoklad normality zamítnut
Uypočtená hladina významnosti : 0.0000E+00
(3) TEST NEZÁVISLOSTI:
Tabulkový kvantil t(1-alfa/2,n+1) : 2.0141E+00
Test autokorelace : 1.6436E+00
Závěr: Předpoklad nezávislosti přijat
Uypočtená hladina významnosti : 5.3616E-02
Předpoklad homogenity výběru:
Napověda-F1 Řádek: 35 - 57 Celken: 85 Délka: 2902
```

```
DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS
U V Y S L E D K Y
(5) DETEKCE ODLEHLÝCH BODŮ:
Bod číslo 6 (horní) : 1.6900E+00
Bod číslo 12 (horní) : 1.8780E+00
Bod číslo 17 (horní) : 3.0100E+00
Bod číslo 24 (horní) : 3.2100E+00
Bod číslo 25 (horní) : 2.5500E+00
Bod číslo 29 (horní) : 1.5100E+00
Počet odlehlých bodů : 6
Parametry s vynechanými odlehlými hodnotami:
Průměr : 1.6818E-01 Rozptyl : 2.8225E-02
Směrodatná odchylka : 1.6800E-01 Šikmost : 1.7755E+00
Špičatost : 6.1925E+00
Protože jde o asymetrické rozdělení, nelze si všimnout detekce odlehlých bodů (outlierů).
Napověda-F1 Řádek: 71 - 85 Celken: 85 Délka: 2902
```

2. kapitola

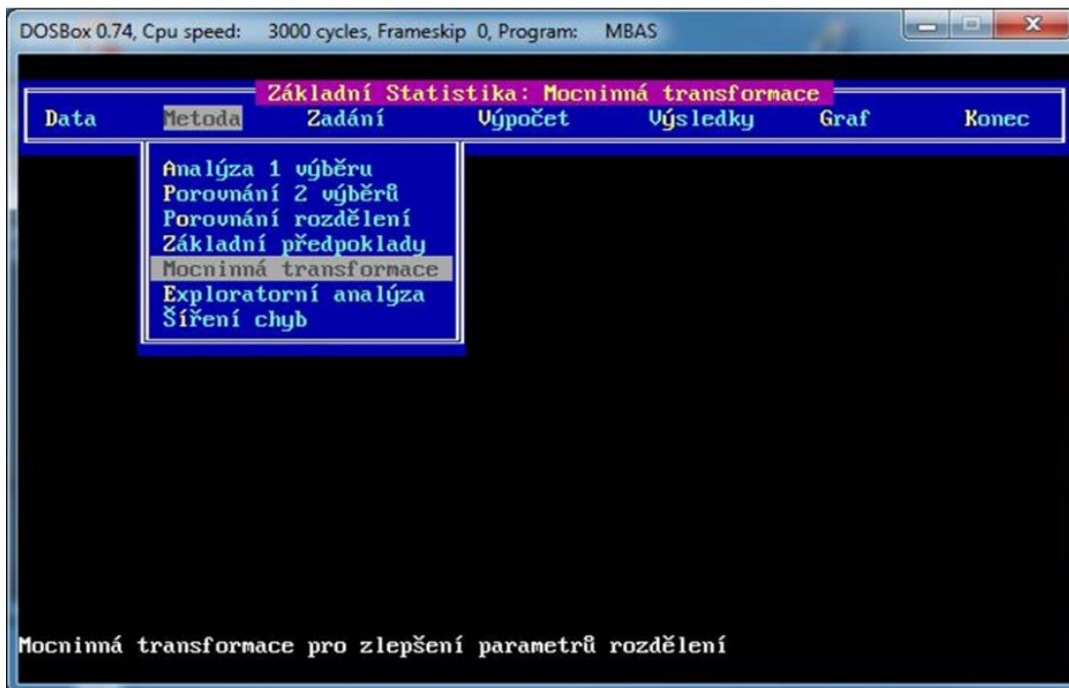
# EDA: Transformace dat

Řešení úloh z Kompendia:

# B204



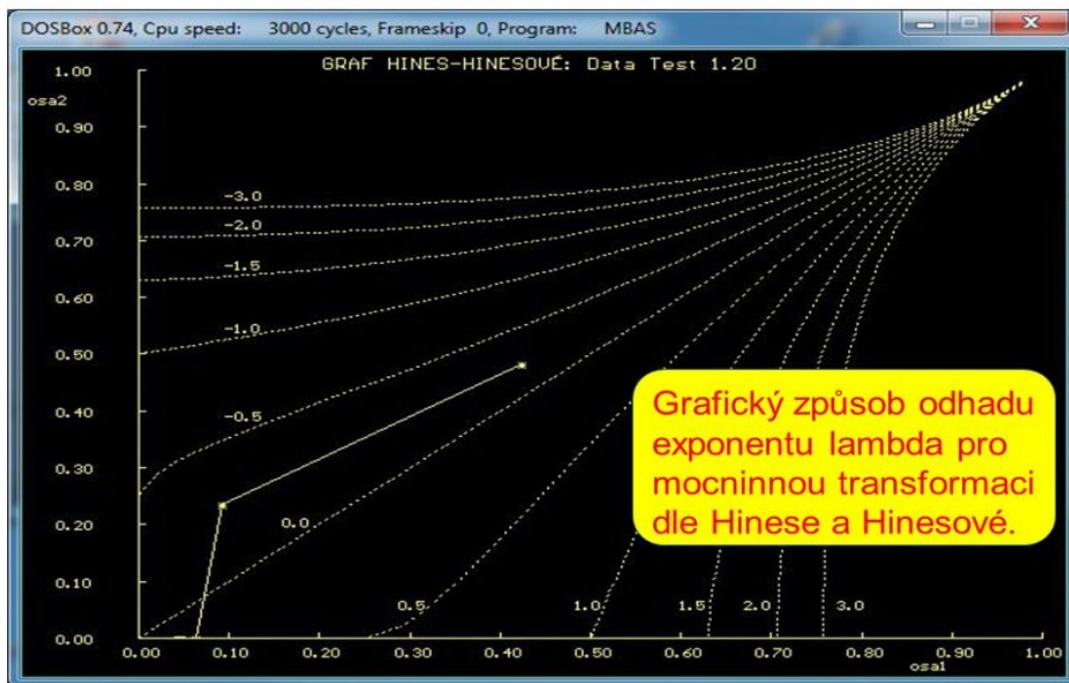
## Volba metody: Mocninná transformace a Box-Coxova transformace



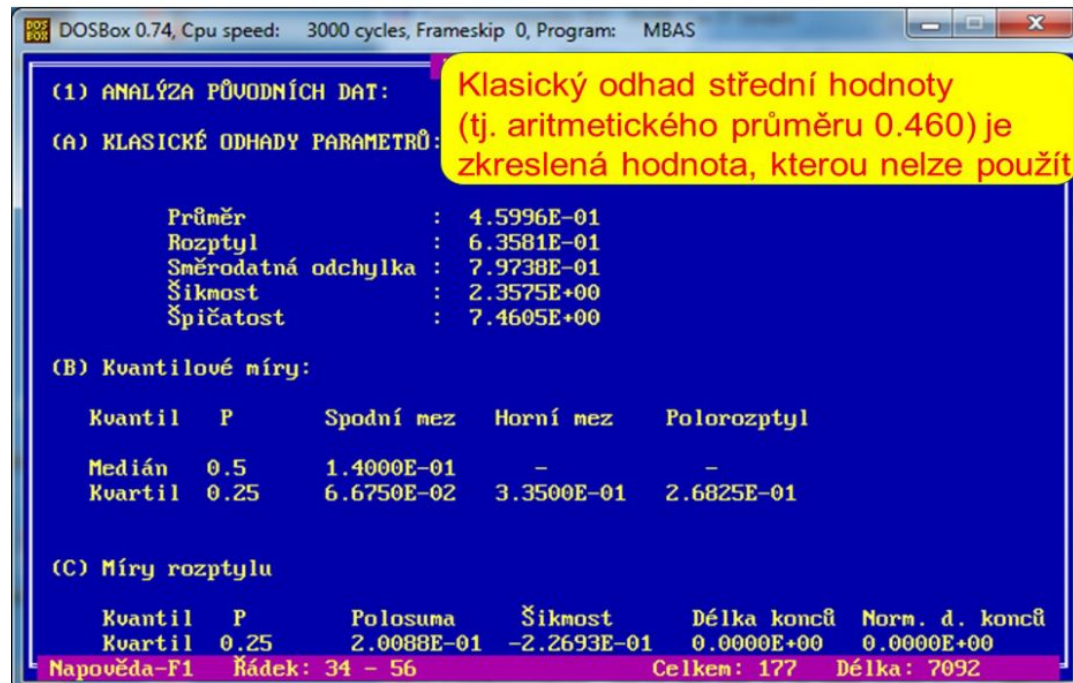
## Diagnostika: Rozhodčí kritérium k užití transformace a odhad mocniny



## 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
U 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
U 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:
Kvantil P
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

```
DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS
U 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
```

Ukázka původních a transformovaných dat po setřídění.

# 2. kapitola EDA: Odhady UDA

Řešení úloh z Kompendia:

# B204

## Volba metody

DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS

Základní Statistika: Analýza jednoho výběru

Data	Metoda	Zadání	Účpočet	Účsledky	Graf	Konec
	Analýza 1 výběru					
	Porovnání 2 výběrů					
	Porovnání rozdělení					
	Základní předpoklady					
	Mocnná transformace					
	Exploratorní analýza					
	Šíření chyb					

Odhady parametrů polohy, rozptýlení a tvaru původních dat, když nebudeme respektovat asymetrii rozdělení. Z výstupu jsou použitelné pouze robustní odhady.

Základní statistické hodnoty, momenty

## Output: Robustní odhady polohy, rozptýlení a tvaru

DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS

U Ý S L E D K Y

```
Průměr : 2.5288E-01
Směr. odchylka : 6.5078E-01
Rozptyl : 4.2351E-01
Průměr, winsor. : 3.7450E-01
St.odch. winsor. : 5.8956E-01
Rozptyl, winsor. : 3.4758E-01
95.0% spolehlivost:
Spodní mez: 5.0769E-02 Horní mez: 4.5498E-01

Uřezání 40% (pro P=0.40):
Průměr : 1.4077E-01
Směr. odchylka : 1.2720E-01
Rozptyl : 1.6179E-02
Průměr, winsor. : 1.4695E-01
St.odch. winsor. : 6.6725E-02
Rozptyl, winsor. : 4.4522E-03
95.0% spolehlivost:
Spodní mez: 9.1215E-02 Horní mez: 1.9033E-01

Biweight:
Průměr : 3.3474E-01
Směr. odchylka : 2.4836E-01
Rozptyl : 6.1685E-02
```

U uřezaných výběrů asymetrického rozdělení ztrácíme informaci, a proto jsou odhady nepoužitelné!!!

Napověda-F1 Řádek: 74 - 96 Celkem: 110 Délka: 3826

## Output: Klasické a robustní odhady polohy, rozptýlení a tvaru

DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS

U Ý S L E D K Y

```
(1) PARAMETRY TVARU:
Šikmost : 2.3575E+00
Špičatost : 7.4605E+00

(2) KLASICKÉ ODHADY PARAMETRŮ :
Průměr : 4.5996E-01
Směr. odchylka : 7.9738E-01
Rozptyl : 6.3581E-01
95.0% spolehlivost:
Spodní mez: 2.1754E-01 Horní mez: 7.0239E-01

(3) OSTATNÍ ODHADY POLOHY:
Odhad modu : 1.4000E-01
Odhad polosumy : 1.6051E+00

(4) ROBUSTNÍ ODHADY PARAMETRŮ :
Medián : 1.4000E-01
Směr. odchylka mediánu: 1.7188E-01
Rozptyl mediánu : 2.9543E-02
Rozptyl (nepar.) : 1.1765E-03
```

Klasické odhady jsou u asymetrického rozdělení nepoužitelné!!!

Robustní odhady jsou u asymetrického rozdělení správně použitelné!

Napověda-F1 Řádek: 33 - 55 Celkem: 110 Délka: 3826

## Output: Adaptivní odhady polohy, rozptýlení a tvaru

DOSBox 0.74, Cpu speed: 3000 cycles, Frameskip 0, Program: MBAS

U Ý S L E D K Y

```
Biweight:
Průměr : 3.3474E-01
Směr. odchylka : 2.4836E-01
Rozptyl : 6.1685E-02
Váhy sqrt(w) : 6.0656E+00
95.0% spolehlivost:
Spodní mez: 2.5217E-01 Horní mez: 4.1732E-01

(5) ADAPTIVNÍ ODHADY PARAMETRŮ:
Hoggovy odhady:
Relativní délka konců : 3.8990E+00
Průměr : 1.4245E-01
Směr. odchylka : 1.6103E-01
Rozptyl : 2.5931E-02
95.0% spolehlivost:
Spodní mez: 9.3496E-02 Horní mez: 1.9141E-01
```

Adaptivní odhady jsou u asymetrického rozdělení použitelné.

Napověda-F1 Řádek: 92 - 110 Celkem: 110 Délka: 3826

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

# Porovnání programů ADSTAT a QCEXPRT

## Při řešení úloh z Kompendia

## 2. kapitola

# EDA: Exploratorní analýza dat

## Řešení úloh z Kompendia:

# B201

### 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-číslíkové schéma formou sedmipísmenného 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 odlehle body?

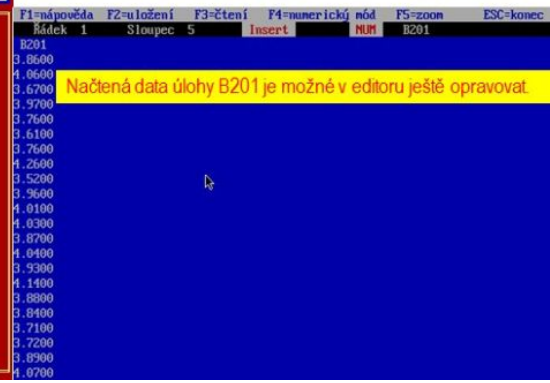
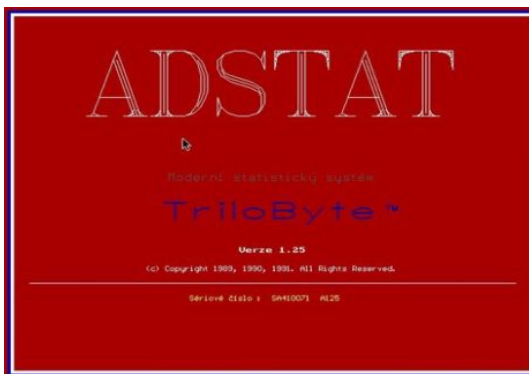
(4) Jaký maximální obsah léčiva v krvi má 75% pacientů?

(5) Odhadněte hloubku prvku 4.00?

**Data:** Obsah léčiva v krvi [mg.l!1]

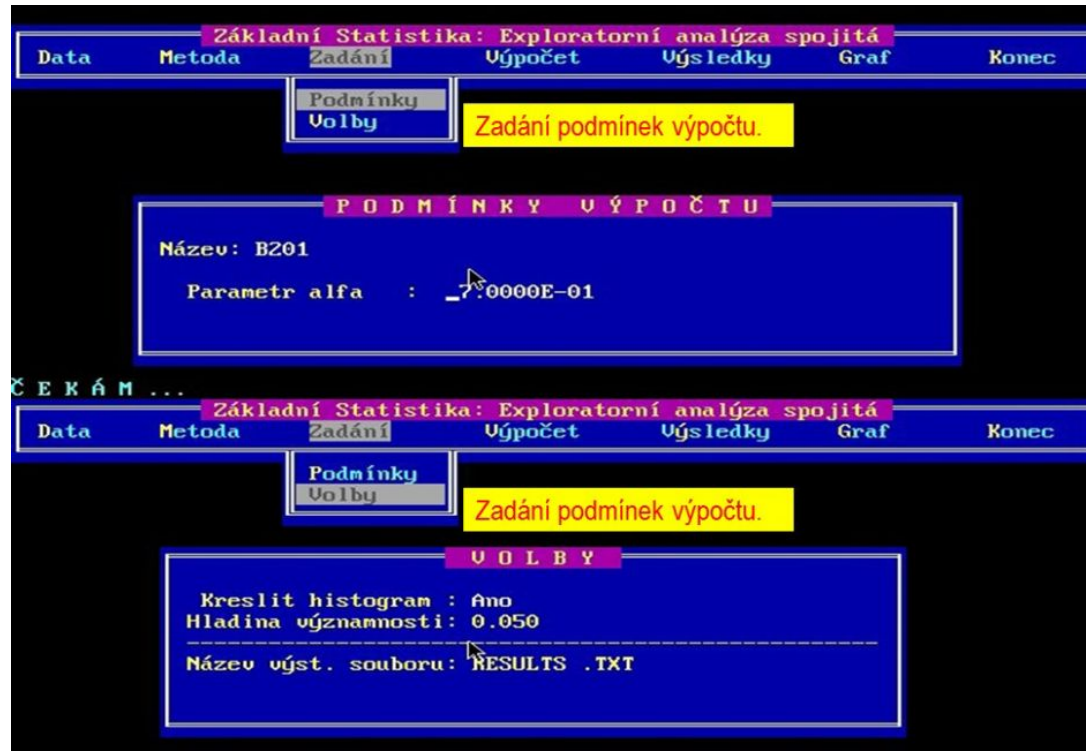
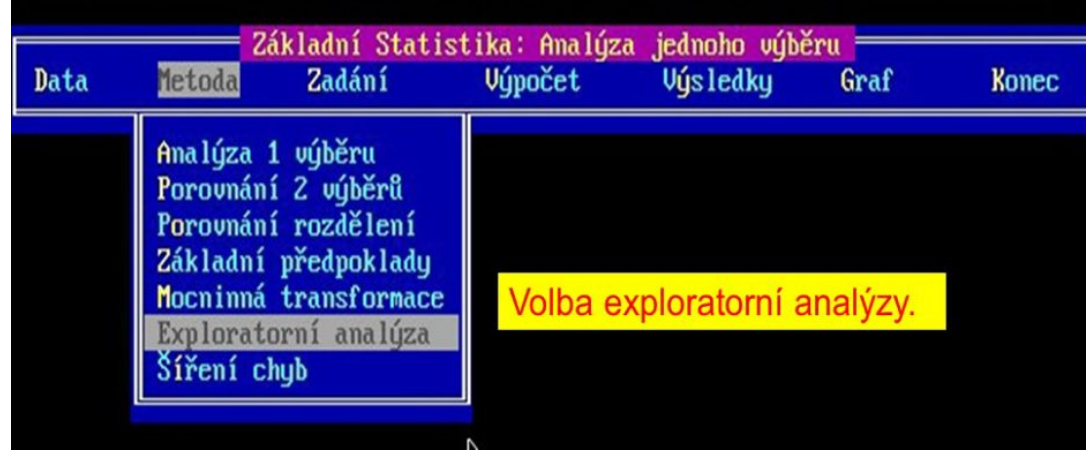
3.86 4.06 3.67 3.97 3.76 3.61 3.76 4.26 3.52 3.96

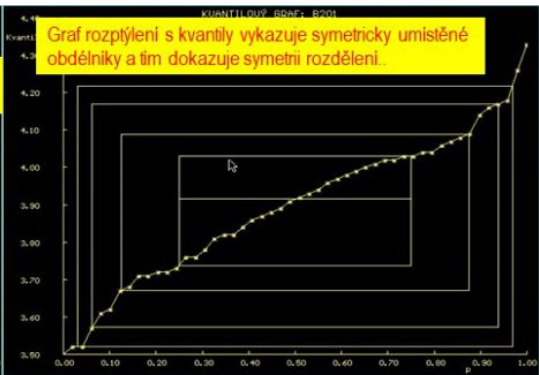
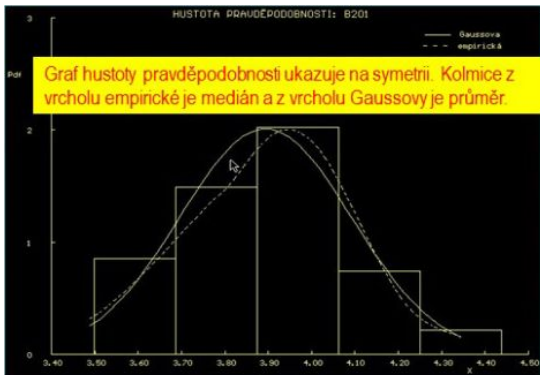
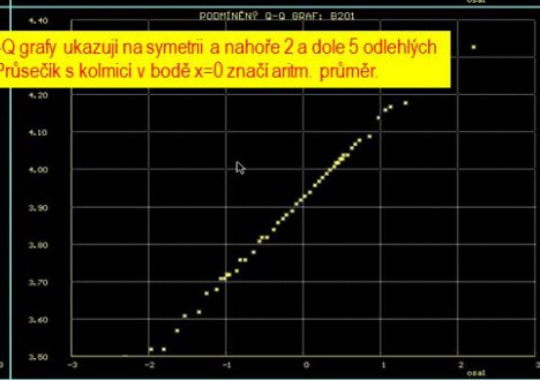
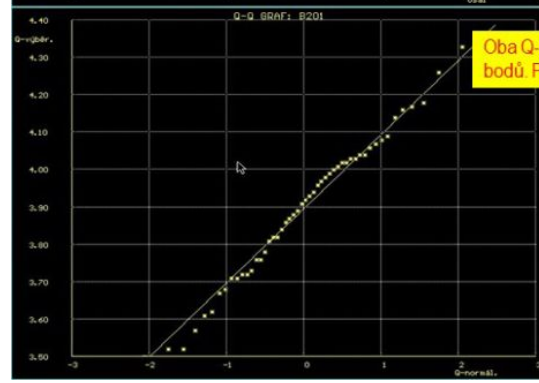
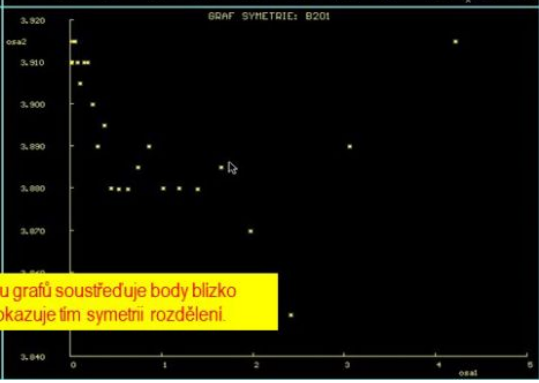
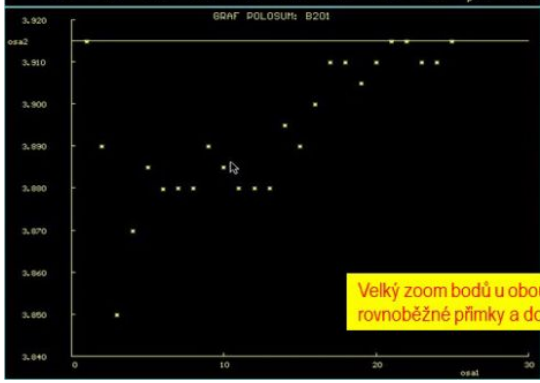
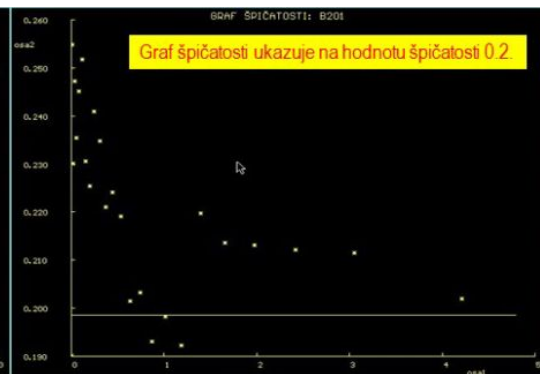
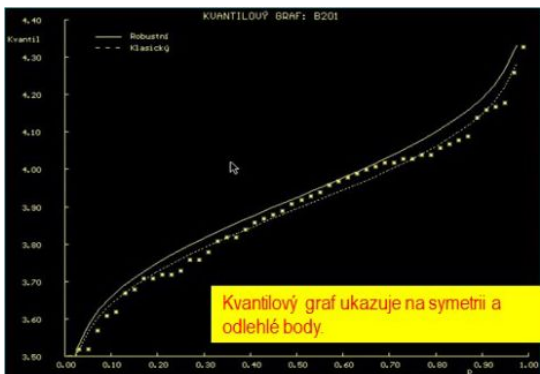
.....



# 1. Průzkumová analýza dat:

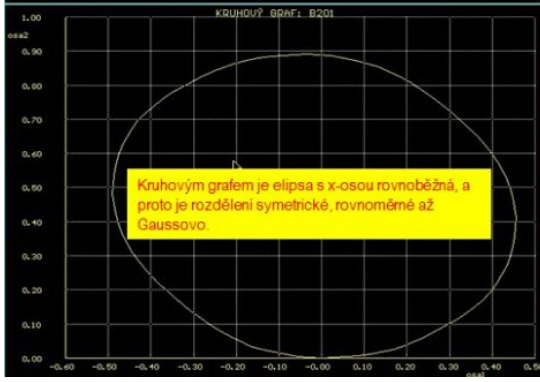
Diagnostické grafy: *stupeň symetrie rozdělení*  
*lokální koncentrace dat*  
*vybočující data*





Použití QC-EXPERTU a načtení matice dat B2 u druhé kapitoly.

	B201	B202	B203
1	3.86	0.8544	49.8
2	4.06	0.8121	47.8
3	3.87	0.8438	50.0
4	3.97	0.8510	50.1
5	3.76	0.8592	50.1
6	3.81	0.8525	50.1
7	3.76	0.8515	50.2
8	4.26	0.8545	50.2
9	3.52	0.8519	50.2
10	3.96	0.8504	50.2
11	4.01	0.8416	50.3
12	4.03	0.8658	50.3
13	3.87	0.8626	55.3
14	4.04	0.8546	50.3
15	3.93	0.8342	50.3
16	4.14	0.8413	50.4
17	3.88	0.8588	50.4
18	3.94	0.8531	50.4
19	3.71	0.8461	50.5
20	3.72	0.8196	50.5
21	3.89	0.8500	50.5
22	4.07	0.8090	6.424
23	3.82	0.8800	0.330
24	4.33	0.2100	4.522
25	4.00	2.5500	2.561
26	3.99	0.2100	1.557
27	4.02	0.4100	1.449
28	3.82	0.0800	4.547
29	3.82	1.5100	0.396
30	3.88	0.1300	1.505



- 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.

Načtení druhé kapitoly.

Volba výpočtů Základní statistika.

Název úlohy: B201 **Zadání výpočtu této úlohy.** Časová osa

Rád trendu: 4 Sloupce: B201, B203, B204, B205, B206, B207a

Testuj hodnotu: 0

Vyhlazení hustoty: 0.5

Rád autokorelace: 4

Hladina významnosti: 0.05

K výpočtu použij: Všechna data, Sloupce, Průměry podskupin

Popis: [Zádný]

Protokoly: Klasické parametry, Trendy vyhlazení, Robustní parametry, Test normality, Vybočující body, Autokorelace, Významnost trendu, Vyhlazení hodnoty, Residua

Standardní, Všechny grafy, Méně

**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.

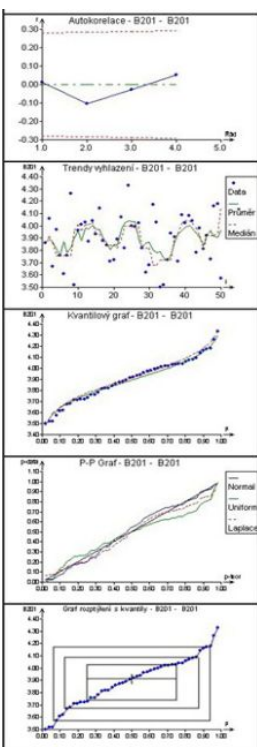
**Sloupce** Výpočet se provede pro každý z vybraných sloupců zvlášť.

**Průměry podskupin** Výpočet se provede pro řádkové průměry z vybraných sloupců. Pokud nejsou sloupce stejné 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.

**Data a parametry:** Data jsou organizována do sloupců. V prvním řádku jsou názvy sloupců. 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 sloupce mít stejnou délku. Minimální počet sloupců je 1. Minimální počet dat je 2.

**Řád trendu** určuje z kolika po sobě jdoucích dat budou počítány klouzavé průměry a klouzavé mediány. Hodnota by měla být menší 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.



**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.

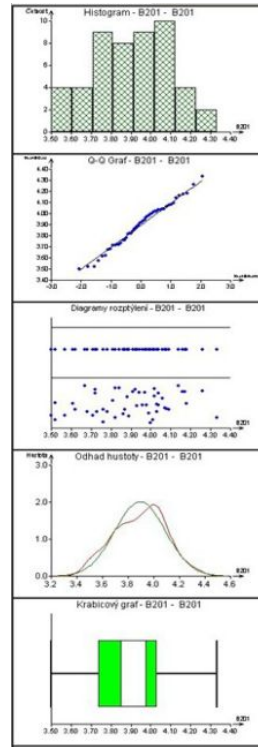
Grafické znázornění trendu v datech pomocí klouzavého průměru (plná křivka) a klouzavého mediánu (přerušovaná křivka) na základě "Řádu trendu" zadaného v dialogovém panelu. Čím je řád trendu větší, tím jsou křivky hladší, méně citlivé na lokální poruchy a zachycují spíše globální dlouhodobý trend. Menší 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 nejbližně 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 rozptýlení dat.



**Grafy výstupu Exploratorní analýzy a jejich vysvětlení.**

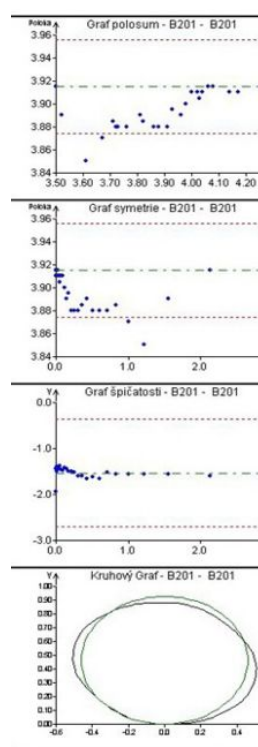
**Histogram:** nejstarší diagram rozdělení ukazuje na symetrii rozdělení. Histogram četností dat v jednotlivých třídách s konstantní šířkou, optimální počet tříd je stanovován automaticky s ohledem na počet dat.

**Q-Q graf:** kvantil-kvantilový graf vykazuje většinu bodů na přímce.

4 body dole a 2 body nahoře jsou podezřelé, že jsou odlehle. Graf pro diagnostiku normality a odlehklých měření, pro normální data bez odlehklých měření má tvar přímky, pro normální data s odlehklými 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 ("rozcházená") 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 rozptění. Porovnání průběhu hustoty pravděpodobnosti normálního rozdělení (plná



**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 níž 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 odchylce 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

## 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

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

## Znaménkový test :

Závěr :	Data jsou nezávislá
---------	---------------------

<b>Znaménkový test</b>	Neparametrický test nezávislosti dat, závislost je indikována, obsahují-li data shluky po sobě jdoucích sekvencí se shodným znaménkem odchylky od průměru.
<b>Závěr</b>	Slovní závěr testu závislosti dat.
<b>Test normality</b>	Kombinovaný test normality založený na shodě šikmosti a špičatosti s normálním rozdělením.

## 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

<b>Normalita</b>	Slovní závěr testu na zadané hladině významnosti.
<b>Vypočtený</b>	Vypočtená testovací statistika.
<b>Teoretický</b>	Příslušný kvantil t-rozdělení.
<b>Pravděpodobnost</b>	Pravděpodobnost odpovídající vypočtené statistice.
<b>Vybočující body</b>	Robustní test na přítomnost vybočujících měření založený na kvantilovém odhadu vnitřních mezí dat.
<b>Homogenita</b>	Slovní závěr testu, nejsou-li v datech vybočující měření, je předpoklad homogenity přijat.
<b>Počet vybočujících bodů</b>	Počet případných měření přesahujících přípustné meze, které je možno považovat za vybočující.
<b>Dolní hranice</b>	Dolní hranice, pod níž je možno data považovat za vybočující.
<b>Horní hranice</b>	Horní hranice, nad níž je možno data považovat za vybočující.

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

## Robustní parametry :

Název sloupce :	B201
Medián :	3.915
IS spodní :	3.709937536
IS horní :	4.120062464
Medianová směr. odchylka :	0.1020426907
Medianový rozptyl :	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

## 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

## Test významnosti trendu :

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

## Robustní parametry :

<b>Medián</b>	Odhad mediánu, tedy 50% kvantilu. Tento odhad střední hodnoty je spolehlivější než aritmetický průměr v případě porušení normality dat nebo přítomnosti vybočujících bodů.
<b>IS spodní</b>	Spodní mez intervalu spolehlivosti mediánu na zadané hladině významnosti.
<b>IS horní</b>	Horní mez intervalu spolehlivosti mediánu na zadané hladině významnosti.
<b>Mediánová směr.odchylka</b>	Odhad směrodatné odchylky na základě mediánu.
<b>Mediánový rozptyl</b>	Odhad rozptylu na základě mediánu.
<b>10% uřezaný průměr</b>	Aritmetický průměr pro symetrickém uřezání 10% dat, tedy 5% nejmenších a 5% největších hodnot. Tento robustní odhad střední hodnoty se doporučuje v případě podezření na vybočující body.
<b>IS spodní</b>	Spodní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
<b>IS horní</b>	Horní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
<b>Rozptyl</b>	Odhad rozptylu na základě mediánu.
<b>Směr. odchylka</b>	Odhad směrodatné odchylky na základě mediánu.
<b>40% uřezaný průměr</b>	Aritmetický průměr pro symetrickém uřezání 40% dat, tedy 20% nejmenších a 20% největších hodnot. Tento robustní odhad střední hodnoty se doporučuje v případě podezření na velký počet vybočujících bodů.
<b>IS spodní</b>	Spodní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
<b>IS horní</b>	Horní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
<b>Rozptyl</b>	Odhad rozptylu na základě mediánu.
<b>Směr. odchylka</b>	Odhad směrodatné odchylky na základě mediánu.

## 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 nevýznamnosti autokorelačního koeficientu; je-li menší než zvolená hladina významnosti, je autokorelace významná.

## R0Krit

Kritická hodnota autokorelačního koeficientu, nad níž se korelace považuje za významnou.

## Výsledek

**Vyhlazené hodnoty** Slovní vyjádření významnosti autokorelace. 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, (nameřená - vyhlazená).

## 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 položené daty

## Pravděpodobnost

Pravděpodobnost 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. Ověření předpokladů výběru dat:

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



# 2. kapitola EDA: Ověření předpokladů o výběru

## Řešení úloh z Kompendia: B201

The image contains two screenshots of statistical software output. The top screenshot shows the 'U V S L E D K Y' (RESULTS) window for 'ZÁKLADNÍ STATISTIKA' and 'Základní předpoklady'. It displays summary statistics and a table of data points. The bottom screenshot shows the 'U V S L E D K Y' window for 'KLASICKÉ ODHADY PARAMETRŮ' (CLASSICAL ESTIMATION OF PARAMETERS). It displays the results of three tests: (1) KLASICKÉ ODHADY PARAMETRŮ, (2) TEST NORMALITY, and (3) TEST NEZÁVISLOSTI. Each test result is highlighted with a yellow box.

**U V S L E D K Y**  
ZÁKLADNÍ STATISTIKA  
Základní předpoklady  
Data a podmínky výpočtu  
Název : B201  
U S T U P  
(1) DATA A PODMÍNKY:  
Počet dat : 50  
Hladina významnosti alfa : 0.050  
Název výstupního souboru : RESULTS.TXT  
(2) USTUPNÍ DATA:  
( 1 ) 3.8600E+00 ( 14 ) 4.0400E+00 ( 27 ) 4.0200E+00 ( 40 ) 4.0200E+00  
( 2 ) 4.0600E+00 ( 15 ) 3.9300E+00 ( 28 ) 3.8200E+00 ( 41 ) 4.0800E+00  
( 3 ) 3.6700E+00 ( 16 ) 4.1400E+00 ( 29 ) 3.6200E+00 ( 42 ) 4.0400E+00  
Napověda-F1 Řádek: 1 - 23 Celkem: 81 Délka: 2739

**U V S L E D K Y**  
aritmetický průměr : 3.8940E+00  
Rozptyl : 3.9398E-02  
Směrodatná odchylka : 1.9849E-01  
Unitární meze:  
Spodní mez : 3.0764E+00  
Horní mez : 4.6837E+00  
(4) MINIMÁLNÍ VELIKOST VÝBĚRU:  
pro 25% relativní chybu směrodatné odchylky : n = 7  
pro 10% relativní chybu směrodatné odchylky : n = 36  
pro 5% relativní chybu směrodatné odchylky : n = 141  
(5) DETEKCE ODHLEHLÝCH BODŮ:  
Ve výběru nejsou odlehlé body  
V datech nejsou outlieri  
Parametry s vymezenými odlehlými hodnotami:  
Průměr : 3.8940E+00 Rozptyl : 3.9398E-02  
Směrodatná odchylka : 1.9849E-01 Šikmost : -1.1915E-01  
Špičatost : 2.4016E+00  
Napověda-F1 Řádek: 60 - 81 Celkem: 81 Délka: 2739

**U V S L E D K Y**  
(1) KLASICKÉ ODHADY PARAMETRŮ:  
Průměr : 3.8940E+00 Rozptyl : 3.9398E-02  
Směrodatná odchylka : 1.9849E-01 Šikmost : -1.1890E-01  
Špičatost : 2.4015E+00  
(2) TEST NORMALITY:  
Tabulkový kvantil  $\chi^2(1-\alpha/2, n-2)$  : 5.9915E+00  
 $\chi^2$ -statistika : 7.8913E-01  
Závěr: Předpoklad normality přijat  
Vypočtená hladina významnosti : 6.7701E-01  
(3) TEST NEZÁVISLOSTI:  
Tabulkový kvantil  $t(1-\alpha/2, n-1)$  : 2.0076E+00  
Test autokorelace : 2.9905E-01  
Závěr: Předpoklad nezávislosti přijat  
Vypočtená hladina významnosti : 3.8306E-01  
Předpoklad homogenity výběru:  
Data jsou homogeni  
Napověda-F1 Řádek: 37 - 59 Celkem: 81 Délka: 2739

### 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.

Základní Statistika: Porovnání rozdělení

Metoda: Zadání Výpočet Výsledky Graf Konec

Podmínky  
Výběr

**V O L B Y**

Kreslit Q-Q graf : Kreslit TDF graf : Ano  
Hladina významnosti: 0.050  
Název výst. souboru: RESULTS .TXT

Určení typu rozdělení výběru.

Data a podmínky výpočtu.

Srovnej výběr s 11 teoretickými rozděleními

Zadej počáteční volby

Základní Statistika: Porovnání rozdělení

Podmínky  
Výběr

**P O D M Í N K Y V Ý P O Č T U**

Název: B201

Počáteční podmínky: Data a podmínky výpočtu.

Parametr pro Weibullovo rozd.: 1.0000E+00  
Parametr pro Paretovo rozděl.: 1.0000E+00  
Parametr pro Gamma rozdělení: 1.0000E+00

Zadej podmínky pro výpočet

**V Ý S L E D K Y**

ZÁKLADNÍ STATISTIKA

Porovnání rozdělení

Název: B201

**D a t a a p o d m í n k y v ý p o č t u**

**V S T U P**

(1) DATA A PODMÍNKY:  
Počet dat : 50  
Hladina významnosti alfa : 0.050  
Název výstupního souboru : RESULTS.TXT

(2) USTUPNÍ DATA:

( 1) 3.8600E+00 ( 14) 4.0400E+00 ( 27) 4.0200E+00 ( 40) 4.0200E+00  
( 2) 4.0600E+00 ( 15) 3.9300E+00 ( 28) 3.8200E+00 ( 41) 4.0900E+00  
( 3) 3.6700E+00 ( 16) 4.1400E+00 ( 29) 3.6200E+00 ( 42) 4.0400E+00  
( 4) 3.9700E+00 ( 17) 3.8800E+00 ( 30) 3.6900E+00 ( 43) 3.7800E+00

Napověďa-F1 Řádek: 1 - 23 Celkem: 158 Délka: 8464

**V Ý S L E D K Y**

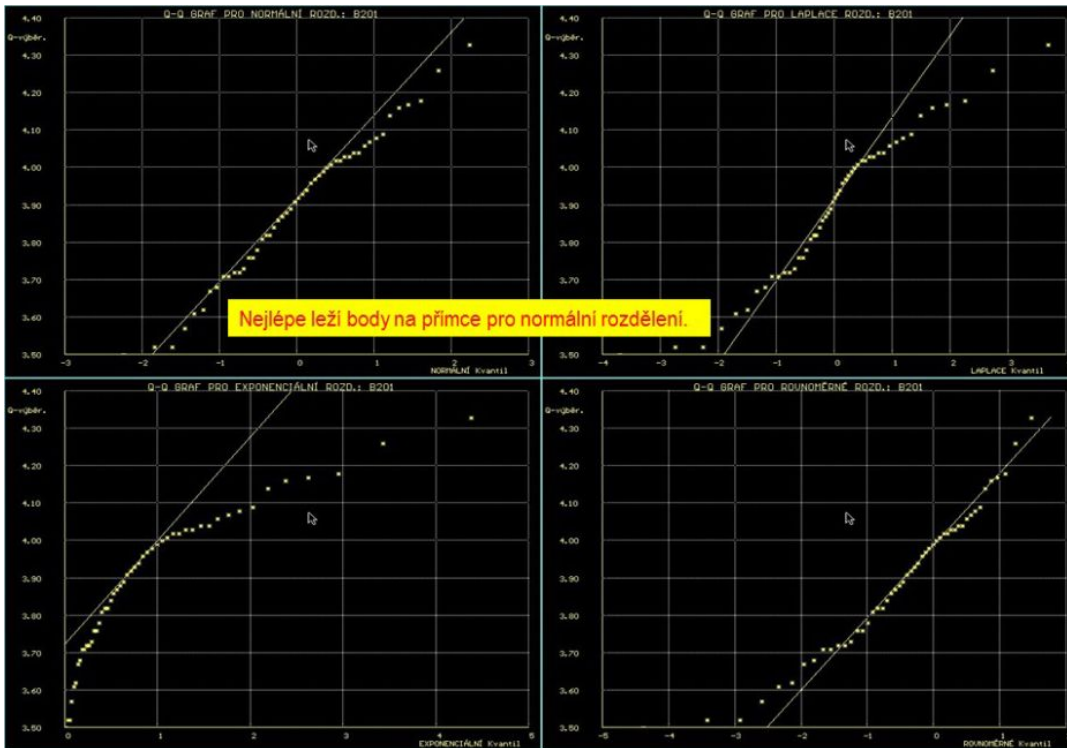
(1) LINEARITA V GRAFU KVANTIL-KVANTIL (Q-Q):

Čís.	Rozdělení	Směrnice	Úsek	Korelační koeficient
0	Laplaceovo	1.4496E-01	3.8940E+00	9.7403E-01
1	Normální	2.0124E-01	3.8940E+00	9.9432E-01
2	Exponenciální	1.8826E-01	3.7087E+00	8.9901E-01
3	Rovnoměrné	6.7467E-01	3.5567E+00	9.8601E-01
4	Lognormální	9.4522E-02	3.7446E+00	8.2336E-01
5	Gumbelova	1.5654E-01	3.9824E+00	9.7451E-01

Sleduje se nejvyšší hodnota korelačního koeficientu blízká 1, a to pro prvních 5 rozdělení.

6Weibullovo( 2.0)	4.2672E-01	3.5171E+00	9.8026E-01
6Weibullovo( 2.5)	5.2410E-01	3.4298E+00	9.8961E-01
6Weibullovo( 3.0)	6.1507E-01	3.3452E+00	9.9358E-01
6Weibullovo( 3.5)	7.0227E-01	3.2623E+00	9.9511E-01
6Weibullovo( 4.0)	7.8709E-01	3.1806E+00	9.9546E-01
6Weibullovo( 4.5)	8.7035E-01	3.0996E+00	9.9523E-01
6Weibullovo( 5.0)	9.5252E-01	3.0191E+00	9.9470E-01
6Weibullovo( 5.5)	1.0339E+00	2.9390E+00	9.9404E-01
6Weibullovo( 6.0)	1.1147E+00	2.8593E+00	9.9332E-01
6Weibullovo( 6.5)	1.1951E+00	2.7798E+00	9.9259E-01

Napověďa-F1 Řádek: 36 - 58 Celkem: 158 Délka: 8464



2. kapitola

**EDA: Transformace dat**

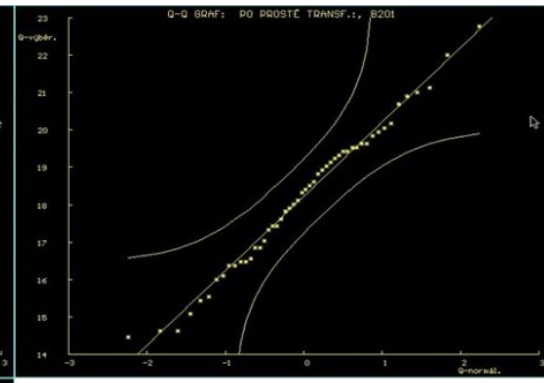
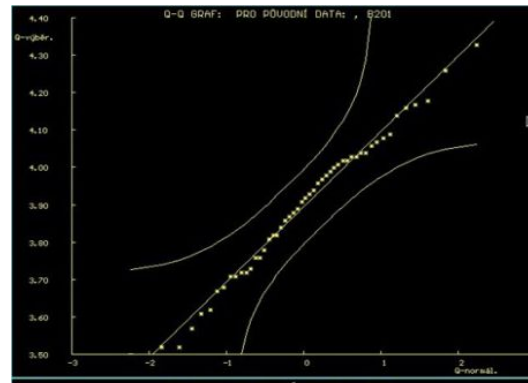
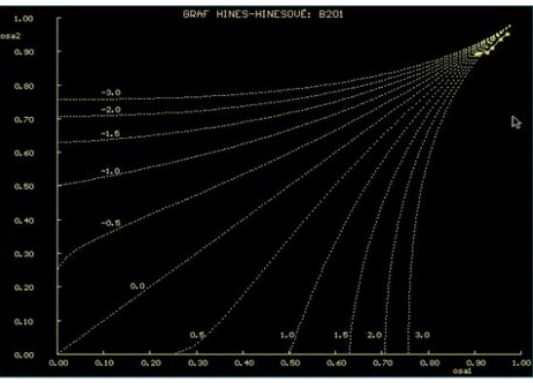
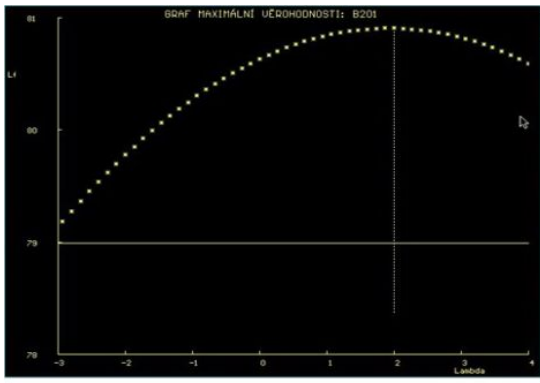
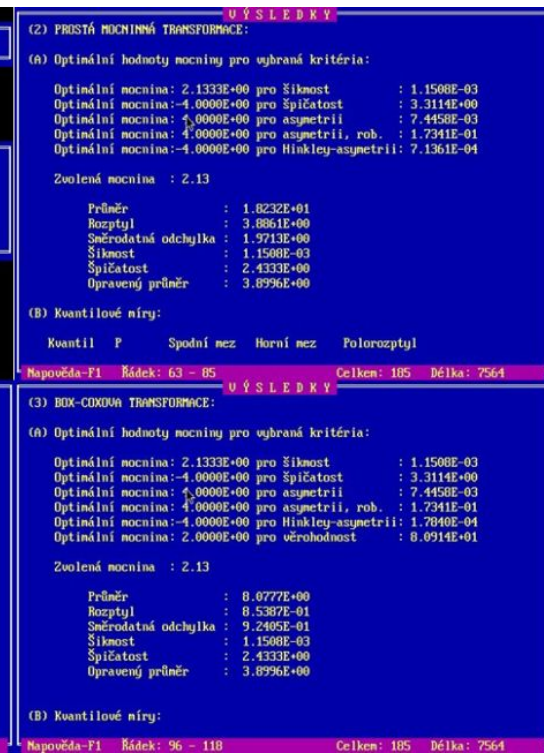
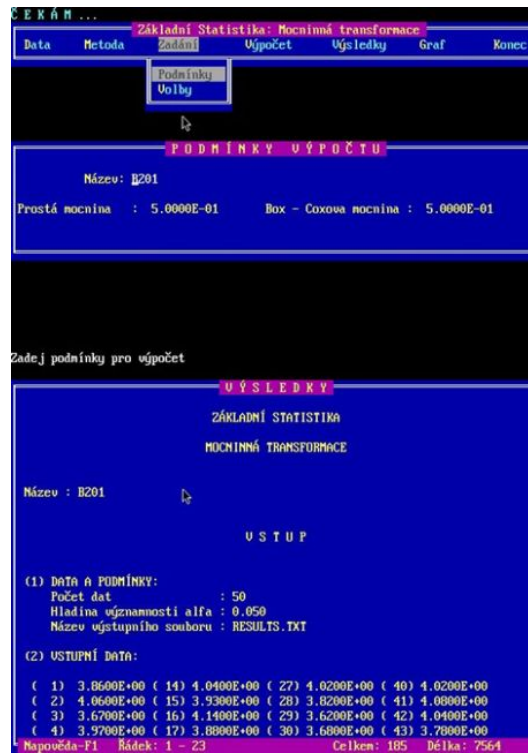
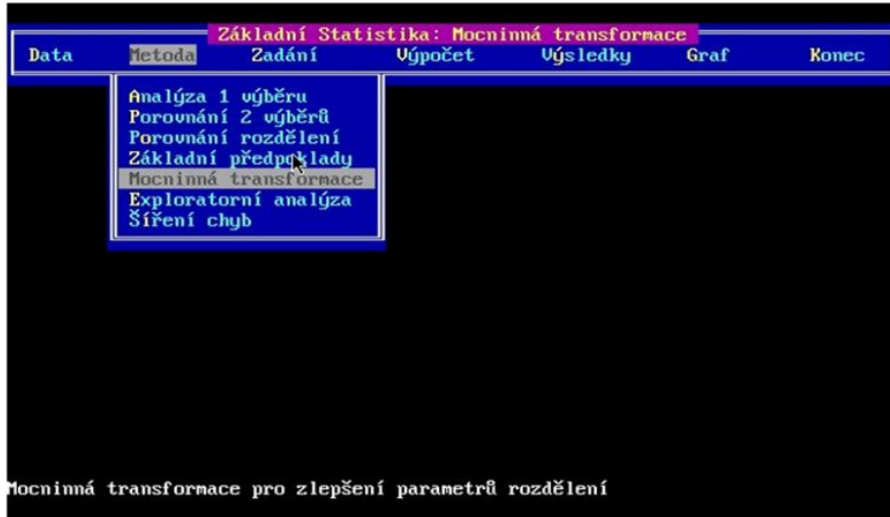
Řešení úloh z Kompendia:

**B201**

# 3. Transformace dat:

Analýza dat:

*originální data*  
*data po mocninné transformaci*  
*data po Box-Coxově transformaci*

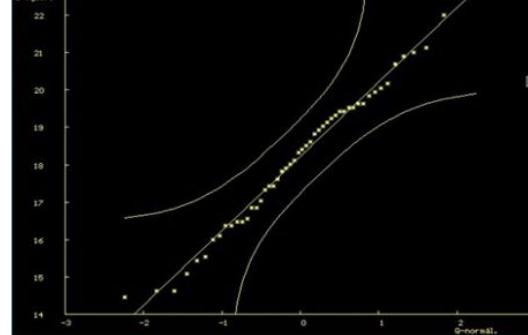


**Rozhodčí kritérium použití transformace dat:**

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

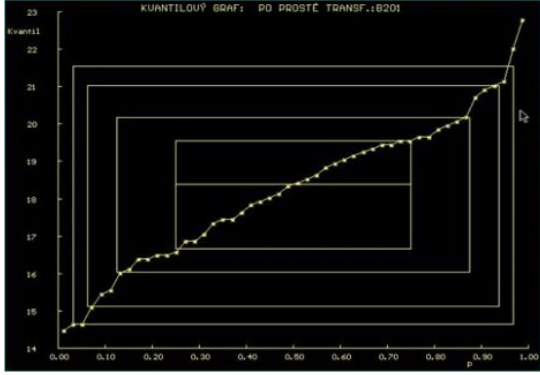
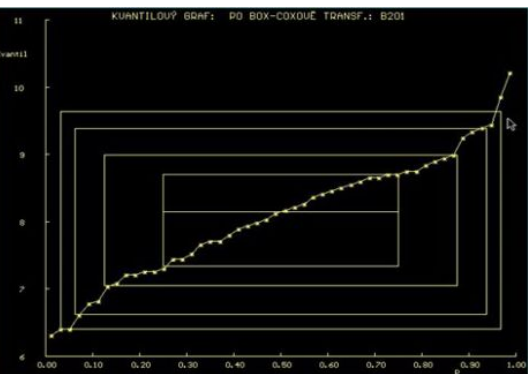
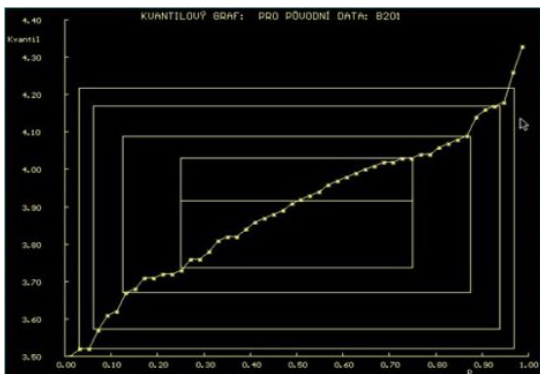
**Určení nejlepšího odhadu exponentu lambda grafickou metodou dle Hinese a Hinesové:**

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.

Název úlohy: B201

Transformace:  Exponenciální  Box-Coxova

Transformovaná data:

Sloupce: B201, B202, B203, B204, B205, B206, B207a, B207b, B208

Data:  Všechna  Označená  Neoznačená  Podle filtru

Popis: [Žádný]

Název úlohy: B201

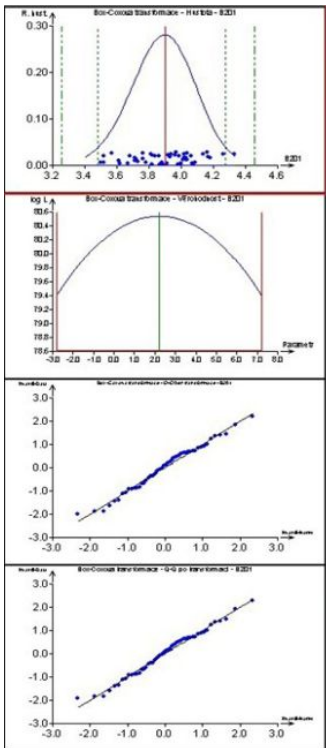
Transformace:  Exponenciální  Box-Coxova

Transformovaná data:

Sloupce: B201, B202, B203, B204, B205, B206, B207a, B207b, B208

Data:  Všechna  Označená  Neoznačená  Podle filtru

Popis: [Žádný]

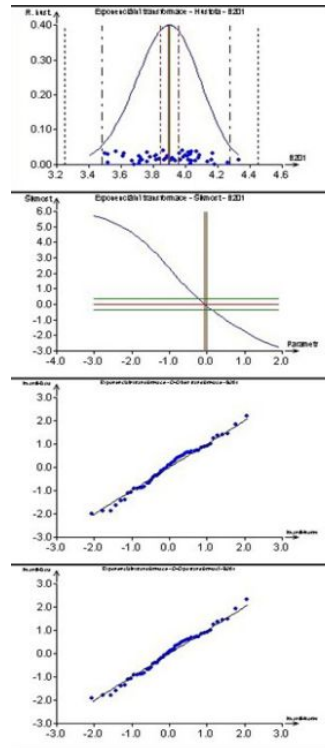


**Graf hustoty** představuje tvar rozdělení, který nejlépe vystihuje data prostředky Box-Coxovy transformace. Svislé čáry představují kvantily (hodnoty) odpovídající (od středu) mediánu (50% kvantil), kvartilu (25% kvantily ohraničující 50% dat),  $\pm 2s$  (zhruba 2.5% kvantily ohraničující interval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

**Graf logaritmu závislosti věrohodnostní funkce** (osa y) na exponentu lambda. Maximu odpovídá optimální hodnota lambda. Vodorovná přímka odpovídá spodní 95% intervalu spolehlivosti maxima věrohodnosti a svislé přímky odpovídají intervalu spolehlivosti odhadu lambda. Obsahovali 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 kvantitativnímu 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),  $\pm 2s$  (zhruba 2.5% kvantily ohraničující interval 95% dat), 0.5% kvantily ohraničující 99% dat a  $\pm 3s$  (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 kvantitativnímu 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 níž 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 spolehlivosti 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. Hodnota uživatelem zadaného parametru v dialogovém panelu. Tato hodnota může být různá od doporučené optimální hodnoty.

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

**Zvolený parametr** 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*

**U V S L E D K Y**

ZÁKLADNÍ STATISTIKA  
 analýza jednorozměrného výběru

Název : B201

U S T U P

(1) DATA A PODMÍNKY:  
 Počet dat : 50  
 Hladina významnosti alfa : 0.050  
 Název výstupního souboru : RESULTS.TXT

(2) USTUPNÍ DATA:

( 1 )	3.8600E+00	( 14 )	4.0400E+00	( 27 )	4.0200E+00	( 40 )	4.0200E+00
( 2 )	4.0600E+00	( 15 )	3.9300E+00	( 28 )	3.8200E+00	( 41 )	4.0800E+00
( 3 )	3.6700E+00	( 16 )	4.1400E+00	( 29 )	3.6200E+00	( 42 )	4.0400E+00
( 4 )	3.9700E+00	( 17 )	3.8800E+00	( 30 )	3.6800E+00	( 43 )	3.7800E+00
( 5 )	3.7600E+00	( 18 )	3.8400E+00	( 31 )	4.1700E+00	( 44 )	3.9800E+00

Napověda-F1 Rádek: 1 - 23 Celken: 112 Délka: 3926

**U V S L E D K Y**

(1) PARAMETRY TVARU:  
 Šikmost : -1.1912E-01  
 Špičatost : 2.4016E+00

(2) KLASICKÉ ODHADY PARAMETRŮ:  
 Průměr : 3.8940E+00  
 Směr. odchylka : 1.9849E-01  
 Rozptyl : 3.9400E-02  
 95.0% spolehlivost:  
 Spodní mez: 3.8376E+00 Horní mez: 3.9504E+00

(3) OSTATNÍ ODHADY POLOHY:  
 Odhad modu : 4.0300E+00  
 Odhad polosuny : 3.9150E+00

(4) ROBUSTNÍ ODHADY PARAMETRŮ:  
 Medián : 3.9150E+00  
 Směr. odchylka mediánu: 2.6014E-01  
 Rozptyl mediánu: 6.7675E-02  
 Rozptyl (nepar.): 2.0250E-03  
 Směr. odchylka mediánu: 4.5000E-02  
 Rozptyl (Marritz): 1.9250E-03  
 Směr. odchylka mediánu: 4.3875E-02  
 95.0% spolehlivost:  
 Spodní mez: 3.8260E+00 Horní mez: 4.0092E+00

Napověda-F1 Rádek: 35 - 57 Celken: 112 Délka: 3926

**U V S L E D K Y**

Uřezání 10% (pro P=0.10):  
 Průměr : 3.8970E+00  
 Směr. odchylka : 2.1447E-01  
 Rozptyl : 4.5999E-02  
 Průměr, winsor. : 3.8936E+00  
 St. odch. winsor. : 1.8942E-01  
 Rozptyl, winsor. : 3.5879E-02  
 95.0% spolehlivost:  
 Spodní mez: 3.8357E+00 Horní mez: 3.9583E+00

Uřezání 40% (pro P=0.40):  
 Průměr : 3.9130E+00  
 Směr. odchylka : 2.7268E-01  
 Rozptyl : 7.4357E-02  
 Průměr, winsor. : 3.9146E+00  
 St. odch. winsor. : 1.1569E-01  
 Rozptyl, winsor. : 1.3304E-02  
 95.0% spolehlivost:  
 Spodní mez: 3.8271E+00 Horní mez: 3.9989E+00

Blueight:  
 Průměr : 3.8966E+00  
 Směr. odchylka : 1.9826E-01

Napověda-F1 Rádek: 75 - 97 Celken: 112 Délka: 3926

**U V S L E D K Y**

(4) ROBUSTNÍ ODHADY PARAMETRŮ:  
 Medián : 3.9150E+00  
 Směr. odchylka mediánu: 2.6014E-01  
 Rozptyl mediánu: 6.7675E-02  
 Rozptyl (nepar.): 2.0250E-03  
 Směr. odchylka mediánu: 4.5000E-02  
 Rozptyl (Marritz): 1.9250E-03  
 Směr. odchylka mediánu: 4.3875E-02  
 95.0% spolehlivost:  
 Spodní mez: 3.8260E+00 Horní mez: 4.0092E+00

Uřezání 5% (pro P=0.05):  
 Průměr : 3.8942E+00  
 Směr. odchylka : 2.0668E-01  
 Rozptyl : 4.2717E-02  
 Průměr, winsor. : 3.8988E+00  
 St. odch. winsor. : 1.9526E-01  
 Rozptyl, winsor. : 3.8128E-02  
 95.0% spolehlivost:  
 Spodní mez: 3.8356E+00 Horní mez: 3.9528E+00

Uřezání 10% (pro P=0.10):  
 Průměr : 3.8970E+00  
 Směr. odchylka : 2.1447E-01  
 Rozptyl : 4.5999E-02  
 Průměr, winsor. : 3.8936E+00  
 St. odch. winsor. : 1.8942E-01  
 Rozptyl, winsor. : 3.5879E-02  
 95.0% spolehlivost:  
 Spodní mez: 3.8357E+00 Horní mez: 3.9583E+00

(5) ADAPTIVNÍ ODHADY PARAMETRŮ:  
 Hoppovy odhady:  
 Relativní délka konců : 2.3753E+00  
 Průměr : 3.8940E+00  
 Směr. odchylka : 1.9849E-01  
 Rozptyl : 3.9400E-02  
 95.0% spolehlivost:  
 Spodní mez: 3.8376E+00 Horní mez: 3.9504E+00

Napověda-F1 Rádek: 94 - 112 Celken: 112 Délka: 3926

## **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.