UNIVERZITA PARDUBICE

Fakulta chemicko-technologická

Katedra analytické chemie

15. LICENČNÍ STUDIUM: „GALILEO“

STATISTICKÉ ZPRACOVÁNÍ DAT

Semestrální práce

STATISTICKÁ ANALÝZA JEDNOROZMĚRNÝCH DAT

20015 – 2017

Vedoucí studia a odborný garant:
Prof. RNDr. Milan Meloun, DrSc.

Vyučující:
Prof. RNDr. Milan Meloun, DrSc.

Autor práce:
ANDRII ZAIKA

OBSAH

Úloha 1: Statistická analýza velkých výběrů

Zadání *Výška stromů ve stejnověkém bukovém kotlíku z inventarizace roku 2015*

Měřením metrem byla za průběžní každoroční inventarizace v roce 2015 získána data pro výšky stromů, vyjádřené v cm a měřené od zemi do nejvyšší větve stromů ve stejnověkém bukovém kotlíku. Nalezněte typ rozdělení pro výběr všech dat, proveďte průzkumovou analýzu spojitých dat (EDA), ověřte předpoklady, rozeberte a vysvětlete diagnostiky.

Data: Výška stromů buku lesního [cm]:

48 43 36 41 44 38 32 35 33 32 38 38 34 35 32 35 38 49 38 33 29 28
36 24 30 34 34 27 34 34 37 41 44 41 22 33 29 24 27 43 33 20 31 30
39 43 38 37 35 38 29 43 41 34 45 44 49 39 34 39 34 34 32 52 27 46
36 39 41 35 39 43 33 33 51 36 24 35 22 25 22 28 28 28 39 34 33 42
44 27 48 28 40 23 34 29 42 36 42

Program: ADSTAT, QC-Expert

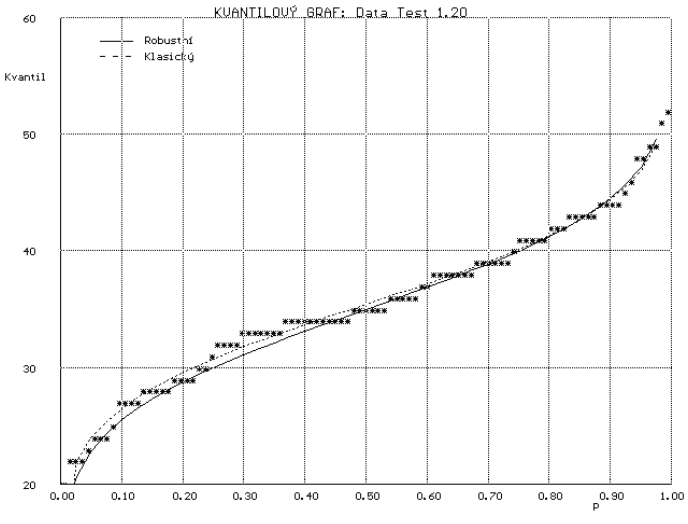
Řešení:

1) jednorozměrná data, exploratorní analýza dat

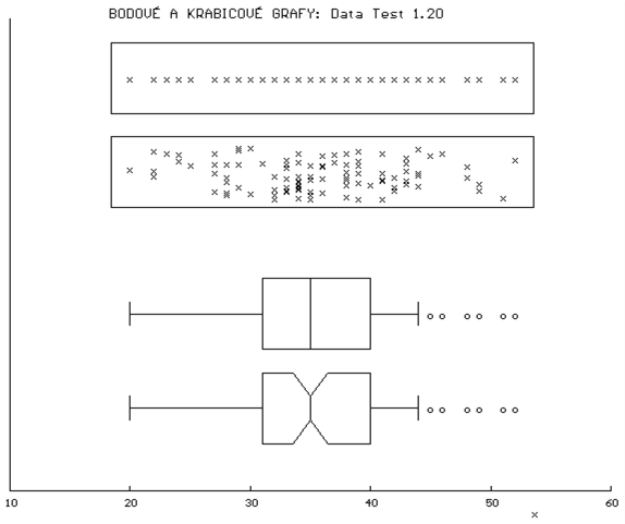
Klasické odhady parametrů:

Medián:	35,00	Průmět:	35,43
Rozptyl:	48,68	Šikmost:	0,045
Špičatost:	2,65	Směrodatná odchylka:	6,98

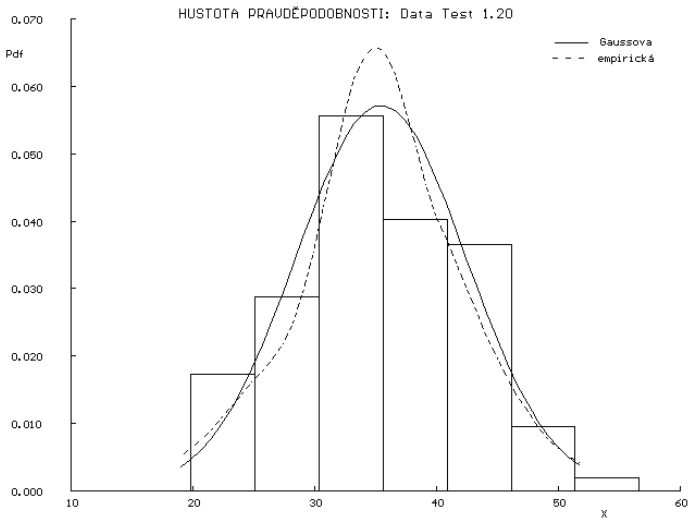
Grafická analýza dat:



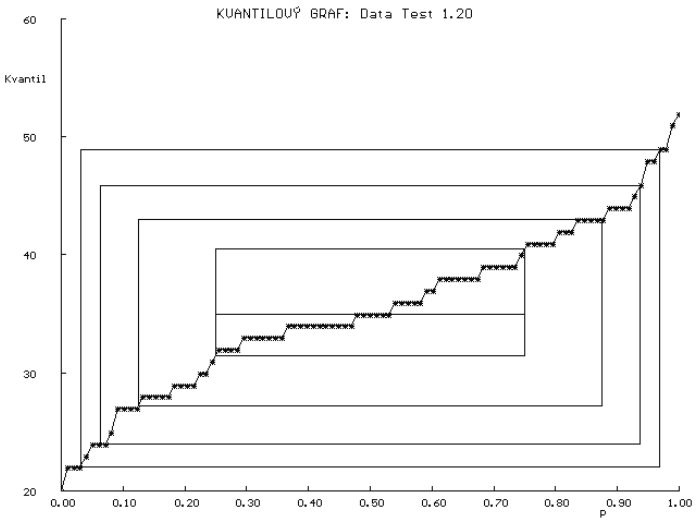
Obr. 1 Kvantilový graf.



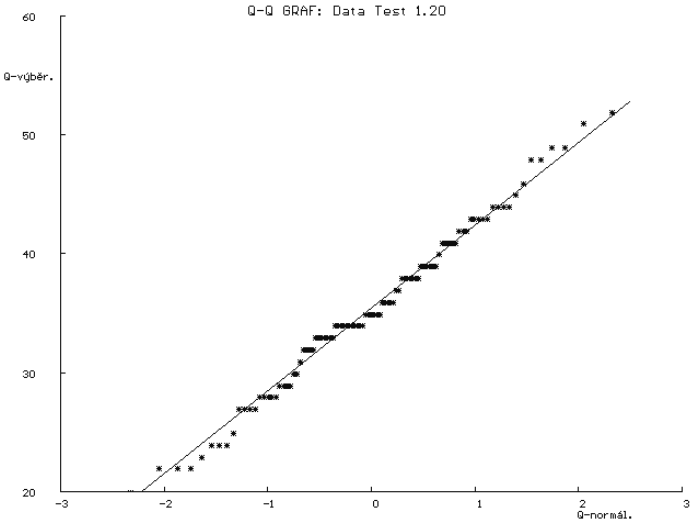
Obr. 2 Diagramy rozptýlení a krabicové grafy.



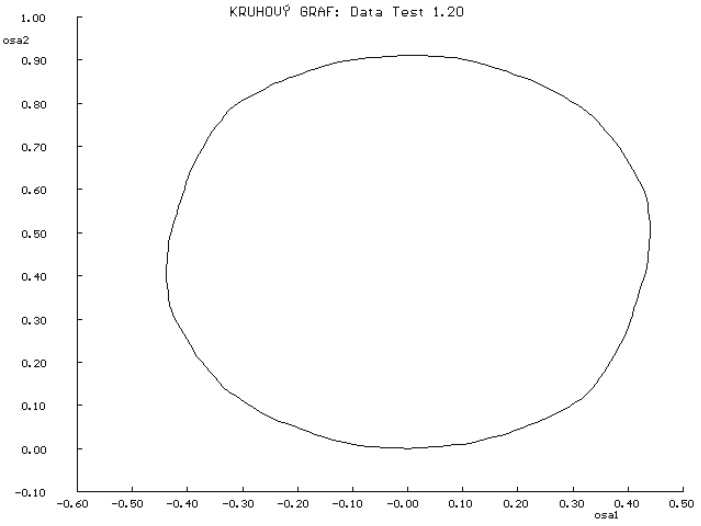
Obr. 3 Graf hustoty pravděpodobnosti.



Obr. 4 Graf rozptýlení s kvantily.



Obr. 5 Kvantil-kvantilový (*Q-Q*) graf.



Obr. 6 Kruhový graf.

Kvantilový graf (Obr. 1) říká, že dvě křivky (klasická a robustní) jsou moc podobné, a tím znamená, že aritmeticky průměr a medián jsou skoro stejní. Krabicový graf (Obr. 2) ukazuje na medián, který rozděluje krabici na dva díly, a v horní části pořádkových statistik lze indikovat 2 až 5 podezřelých bodů. To ještě ale musí potvrdit základní předpoklady pro detekce odlehlých bodů. Linearita kvantil-kvantilovém Q-Q grafu (Obr. 5) ukazuje, že body leží na přímce, a to znamená, že korelační koeficient pro normální rozděleni dosahuje nejvyšší hodnoty. To nám a i potvrdí „porovnání rozdělení“ v ADSTATu, kde r_{xy} pro normální rozdělení je 0,995:

Čís.	Rozdělení	Směrnice	Úsek	Korelační koeficient
0	Laplaceovo	5.0049E+00	3.5434E+01	9.7876E-01
1	Normální	7.0242E+00	3.5434E+01	9.9531E-01
2	Exponenciální	6.5861E+00	2.8907E+01	9.1322E-01
3	Rovnoměrné	2.3616E+01	2.3626E+01	9.7959E-01
4	Lognormální	3.1281E+00	3.0411E+01	8.2577E-01
5	Gumbelovo	5.3750E+00	3.8499E+01	9.6571E-01

Graf hustoty pravděpodobnosti (Obr. 3) dokazuje, že vrchol obou křivek na *x*-ve ose mají skoro stejní hodnoty, to čísla aritmetického průměru a mediánu jsou blízko sebe, tím se statisticky považuje za shodné a totožné. Graf vykazuje, že data mají skoro Gaussovo rozdělení. Graf rozptýlení s kvantily a kruhový graf ukazují na symetrické normální rozdělení, to znamená, že aritmetický průměr se bude moci použit.

Závěr EDA: symetrické normální rozdělení, M = 35,00 $\bar{X}$ = 35,43 – lze použit, šikmost g_1 = 0,045 je v rozmezí -0.3 < g_1 < 0.3 a muže se považovat za 0, špičatost g_2 = 2,65 je mezi Gaussovým a rovnoměrným rozdělením, ale leží v rozmezí 2 < g_2 < 4, tak muže se považovat za Gaussovo rozdělení.

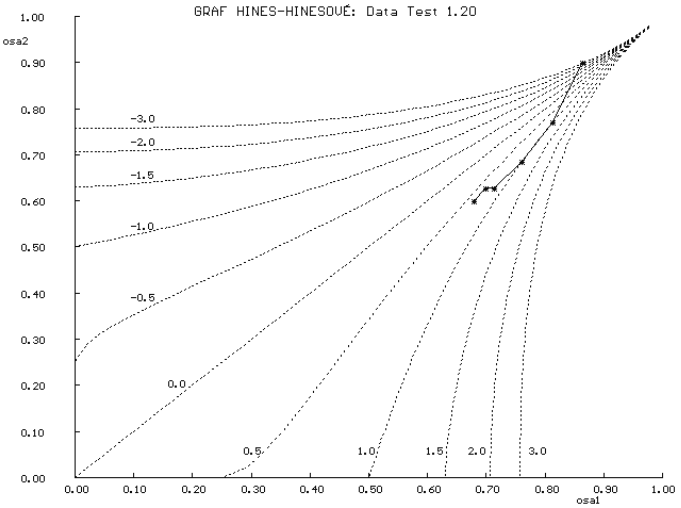
2) Základní předpoklady výběru (ADSTAT)

(a) Test normality: tabulkový kvantil $\chi^2_{1-\alpha}(2)$:	5.992
Odhad statistiky χ^2_{exp}:	0.437
Závěr: Předpoklad normality přijat na spočtené hladině významnosti α = 0.804	
(b) Test nezávislosti: tabulkový kvantil $t_{1-\alpha/2}(n+1)$:	1.984
Odhad von Neumannovy statistiky t_n:	1.2003
Závěr: Předpoklad nezávislosti přijat na spočtené hladině významnosti α = 0.116	
(c) Detekce odlehlých bodů: metodou modifikované vnitřní hradby	
Závěr: Ve výběru nejsou odlehlé body.	

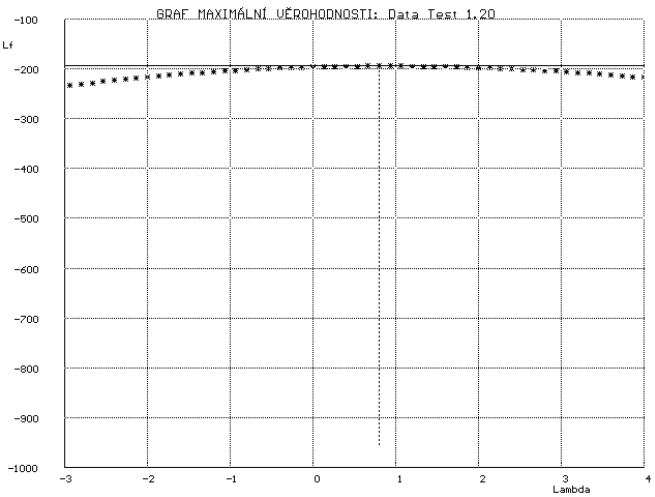
Předpoklad o normalitě je přijat, protože hodnota testovacího kritéria χ^2 je nižší než tabulkový kvantil $\chi^2_{1-\alpha}(2)$. Předpoklad nezávislosti je přijat, protože hodnota testovacího kritéria je nižší než tabulkový kvantil. Data jsou homogenní.

Závěr předpokladů: muže se použit aritmetický průměr.

Pro jistotu svých závěrů je provedena transformace dat.



Obr. 7 Hinesův-Hinesové selekční graf.



Obr. 8 Graf logaritmu věrohodnostní funkce na λ

Pro odhad parametru λ byly použité grafy Hinesův-Hinesové selekční graf (Obr. 7) pro prostu mocninnu transformace, a graf logaritmu věrohodnostní funkce (Obr. 8) pro Box-Coxovu transformace dat. Obě metody k určení optimálního λ ukazuji hodnotu λ = 0,86. Tím, že číslo 1 se nachází nad segmentem na x-ové ose, tak transformace dat bylo možní provést, ale ne moc důležitá, jelikož číslo 1 leží blízko segmentu.

3) Transformace dat

Prosta mocninná transformace dat

Zvolená mocnina:	0,86
Směrodatná odchylka:	3,65
Šikmost:	-0,0238
Špičatost:	2,65
Opravený průměr:	35,34

Box-Coxova transformace

Zvolená mocnina:	0,86
Směrodatná odchylka:	4,24
Šikmost:	-0,0238
Špičatost:	2,65
Opravený průměr:	35,34

Závěr transformace dat: výběr pochází z normálního rozdělení, a proto transformaci dat dělat není potřeba. Však po transformace dat se dostalo k nejlepšímu odhadu střední hodnoty s menši šikmosti a směrodatné odchylky.

4) Odhady parametrů

Analýza jednorozměrného výběru dat

Parametry tvaru:

Šikmost	0,0453
Špičatost	2,65

Klasické odhady parametrů:
Poznámka: Lze použít pro normální rozdělení dat!

Průměr	35,43
Směrodatná odchylka průměru	6,98
Rozptyl průměru	48,68
95 % spolehlivost:	
Spodní mez	34,04
Horní mez	36,83

Robustní odhady parametrů:

Medián	35,00
Směrodatná odchylka mediánu	9,29
Rozptyl mediánu	86,32
95 % spolehlivost:	
Spodní mez	33,27
Horní mez	36,73

Závěr odhadů polohy, rozptylu a tvaru rozdělení: Výše uvedená data zjištěná měřením výšky stromů ve stejnověkém bukovém kotlíku pocházejí z rozdělení normálního. Pro zjištěné **normální rozdělení** dat je použitelné oboje odhady polohy – odhad aritmetického průměru a mediánu, protože jsou skoro stejný. Je možný použít odhad polohy retransformovaného průměru po prosté mocninné nebo Box-Coxové transformaci.

Z úlohy vyplývá, že střední výška stromů ve stejnověkém bukovém kotlíku s 95% pravděpodobnosti je 35,43 cm, směrodatná odchylka je 6,98 cm, spodní konfidenční mez je 34,04 cm a horní konfidenční mez je 36,83 cm.

Úloha 2: Statistická analýza malých výběrů dle Horna

Zadání: Každý měsíc za rok 2013 bylo změřeno sumu srážek na volné ploše ekosystémové stanice v Němčicích. Aplikujte Hornovu metodu pivotů k určení parametrů polohy a rozptýlení a výsledky porovnejte s klasickými a robustními odhady a rozptýlení pomoci software ADSTAT.

Data: Měsíční srážky na volné ploše za rok 2013 [mm]:

34,3	39,6	16,0	30,2	100,8	139,7	3,0	50,5	25,7	46,2	25,4	13,2

Program: EXCELL, ADSTAT

Řešení:

1) Hornův postup

Pořádková statistika:

i	1	2	3	4	5	6	7	8	9	10	11	12
$x_{(i)}$	3,0	13,2	16,0	25,4	25,7	30,2	34,3	39,6	46,2	50,5	100,8	139,7

- *Hloubka pivotu H*

  $H = (\text{int}((n+1)/2)+1)/2 = (\text{int}((12+1)/2)+1)/2 = 3$

- *Dolní pivot x_D, horní pivot x_H*

  $x_D = x_{(H)} = x_{(3)} = 16,0$

  $x_H = x_{(n+1-H)} = x_{(12+1-3)} = x_{(10)} = 50,5$

- *Pivotová polosuma P_L (odhad parametru polohy)*

  $P_L = (x_D + x_H)/2 = (16,0 + 50,5)/2 = 33,3$

- *Pivotové rozpětí R_L (odhad parametru rozptýlení)*

  $R_L = x_H - x_D = 50,5 - 16,0 = 34,5$

- *95%-ní interval spolehlivosti střední hodnoty*

  $P_L - R_L t_{L,0,975}(n) \leq \mu \leq P_L + R_L t_{L,0,975}(n)$

  $33,3 - 34,5 * 0,483 \leq \mu \leq 33,3 + 34,5 * 0,483$

  $16,6 \leq \mu \leq 50,0$

Závěr: Bodový odhad míry polohy je 33,3, míry rozptýlení je 34,5 a intervalový odhad míry polohy je $16,6 \leq \mu \leq 50,0$

2) Analýza jednorozměrného výběru

(a) Test normality: tabulkový kvantil $\chi^2_{1-\alpha}(2)$:	5.992
Odhad statistiky χ^2_{exp}:	11.987
Závěr: Předpoklad normality zamítnut na spočtené hladině významnosti $\alpha = 0.0025$	
(b) Test nezávislosti: tabulkový kvantil $t_{1-\alpha/2}(n+1)$:	2.160
Odhad von Neumannovy statistiky t_n:	0.4534
Závěr: Předpoklad nezávislosti přijat na spočtené hladině významnosti $\alpha = 0.329$	
(c) Detekce odlehlých bodů:	
Závěr: Ve výběru je 1 odlehlý bod, a to bod č. 6 (horní) o hodnotě 139.7.	
(d) Opravené parametry výběru s vynechanými odlehlými hodnotami:	
Odhad aritmetického průměru $\bar{x}$:	34.99
Odhad směrodatné odchylky s:	26.04
Odhad šikmosti $\hat{g}_1$:	1.48
Odhad špičatosti $\hat{g}_2$:	5.29
(e) Prostá mocninná transformace:	
Odhad optimálního exponentu λ	0.27
Odhad průměru transformovaných dat $\bar{y}$	2.56
Opravený odhad průměru původních dat $\bar{x}_R$	33.73
(f) Boxova-Coxova transformace:	
Odhad optimálního exponentu λ	0.27
Odhad průměru transformovaných dat $\bar{y}$	5.83
Opravený odhad průměru původních dat $\bar{x}_R$	33.73
(g) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$:	43.72
Odhad směrodatné odchylky s:	39.11
Odhad šikmosti $\hat{g}_1$:	1.49
Odhad špičatosti $\hat{g}_2$:	4.24
Dolní mez 95.0% intervalu spolehlivosti L_D	18.87
Horní mez 95.0% intervalu spolehlivosti L_H	68.57
(h) Robustní odhady parametrů:	
Medián $\tilde{x}_{0.5}$	32.25
Odhad směrodatné odchylky mediánu $s(\tilde{x}_{0.5})$	28.06
Odhad 40% uřezaného průměru $\bar{x}(40\%)$	32.32
Odhad směrodatné odchylky $s(40\%)$	18.92
Dolní mez 95.0% intervalu spolehlivosti L_D	14.54
Horní mez 95.0% intervalu spolehlivosti L_H	50.09
Odhad M-odhadu střední hodnoty μ_M	36.30
Odhad směrodatné odchylky s_M	33.96
Dolní mez 95.0% intervalu spolehlivosti L_D	13.52
Horní mez 95.0% intervalu spolehlivosti L_H	59.07

3) Porovnání výsledků dle Horna a výsledků ADSTAT

Parametry	Polohy	Rozptýlení	95% interval spolehlivosti střední hodnoty	
Hornův postup	$P_L = 33.3$	$R_L = 34.5$	$L_D = 16.6$	$L_H = 50.0$
Klasický odhad	$x = 43.72$	$s = 39.11$	$L_D = 18.87$	$L_H = 68.57$
Robustní odhad	$x_{0.5} = 32.25$	$s = 28.06$	$L_D = 14.54$	$L_H = 50.09$

Závěr: Vzhledem k počtu analýz referenčního materiálu (n = 12) jsou tedy nejlepšími odhady polohy a rozptýlení výsledky dle Hornova postupu, který je vhodný pro malé výběry ($4 \leq n \leq 20$). Bodový odhad polohy je 33.3 mm, bodový odhad rozptýlení je 34.5 mm, a intervalový odhad polohy je $16{,}6 \leq \mu \leq 50{,}0$.

Úloha 3: Statistické testování

(a) Test správnosti:

Zadání: Výrobce standardního roztoku deklaruje koncentraci diclazurilu 500 μg/ml. Obsah diclazurilu se stanoví vysokoúčinnou kapalinovou chromatografií (HPLC) na reverzní fázi s ternárním gradientem s UV detekcí. Jsou naměřené výsledky správné?

Data: Koncentrace diclazurilu ve standardním roztoku [μ g/ml]:

481	500	511	510	488	520	463	480	532	510

Program: ADSTAT

Řešení: **analýza jednorozměrných dat koncentrace diclazurilu**

Z exploratorní analýzy byla zjištěna asymetrie dat, posun k vyšším hodnotám, v horní části řady pořádkových statistik jeden podezřelý bod. Ze základních předpokladů vyplývá, že předpoklad normality a nezávislosti přijat, soubor neobsahuje odlehlé hodnoty.

Parametry tvaru:

Šikmost	-0,21
Špičatost	2,10

Klasické odhady parametrů:
Poznámka: Lze použít pro normální rozdělení dat!

Průměr	499,50
Směrodatná odchylka průměru	21,125
95 % spolehlivost:	
Spodní mez	484,39
Horní mez	514,61

Robustní odhady parametrů:

Medián	505,00
Směrodatná odchylka mediánu	29,731
95 % spolehlivost:	
Spodní mez	480,80
Horní mez	529,20

Závěr: Pro 95% statistickou jistotu byly tyto následující intervalové odhady: pro aritmetický průměr $\bar{X}$ je interval $484{,}39 \leq \mu \leq 514{,}61$, pro medián $\tilde{x}_{0.5}$ pak $480{,}80 \leq \mu \leq 529{,}20$. Z uvedených intervalových odhadů vyplývá, že obsah diclazurilu 500 μg/ml leží v rozmezí zadané normy a naměřené výsledky jsou správné.

(b) Test shodnosti:

Zadání: Pro stanovení střední hodnoty průměru smrku měřením na kmeni 1.3 nad zemí stromů byly použity dvě metody – Fluryho průměrka a klasickým obvodem metrem v přepočtu na tloušťku. Vedou obě metody ke stejným výsledkům?

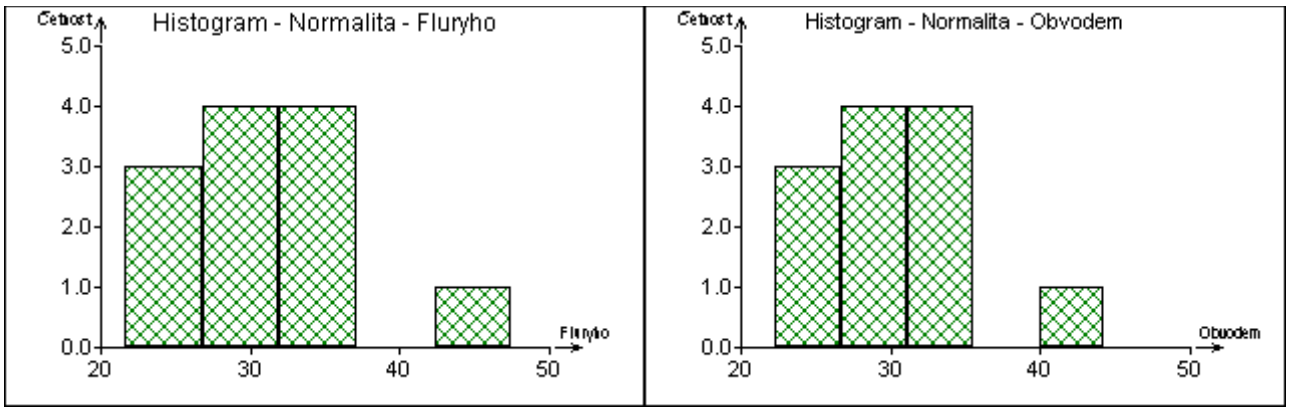
Data: Obvod stromů smrku ztepilého [cm]:

21,6	23,0	31,1	32,2	31,1	30,5	32,2	35,1	47,5	34,9	29,6	24,8

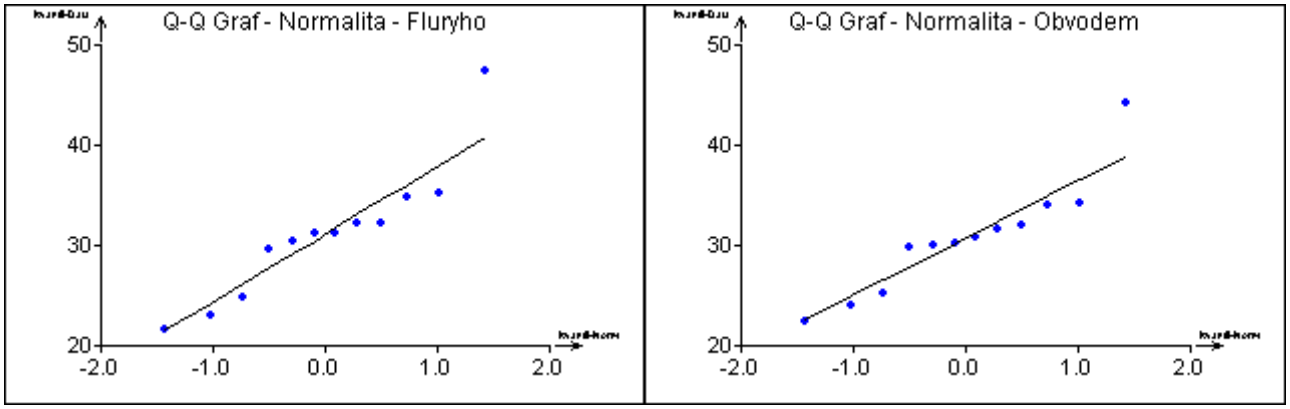
22,3	24,1	30,2	34,2	30,9	29,8	32,1	34,0	44,3	31,7	30,0	25,2

Program: QC Expert

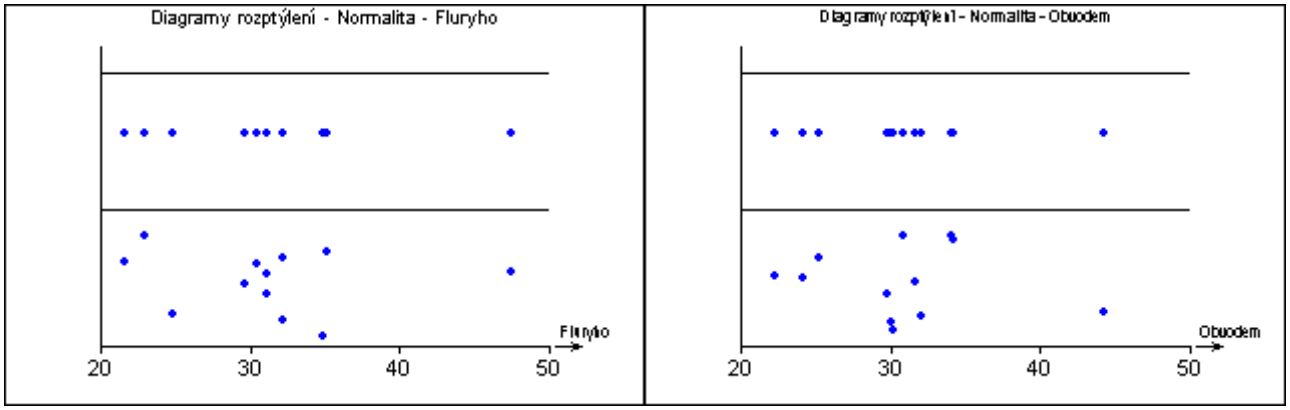
***Řešení:* základní statistika, porovnání dvou výběrů**



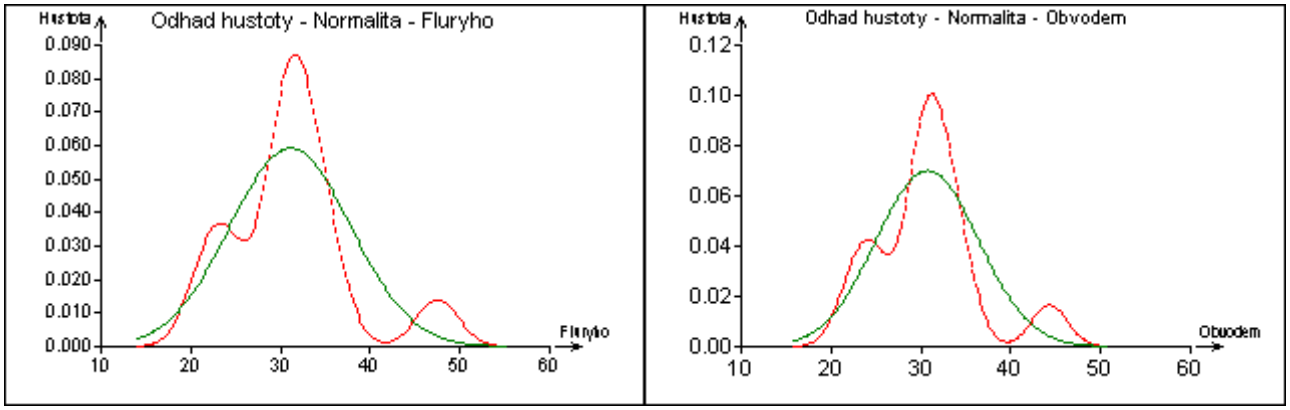
Obr. 9 Histogram



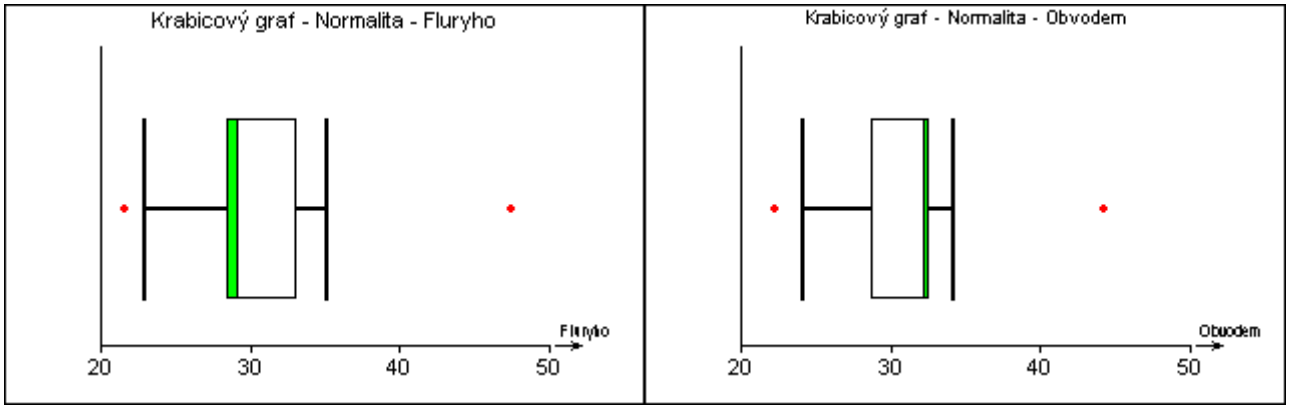
Obr. 10 Q-Q graf



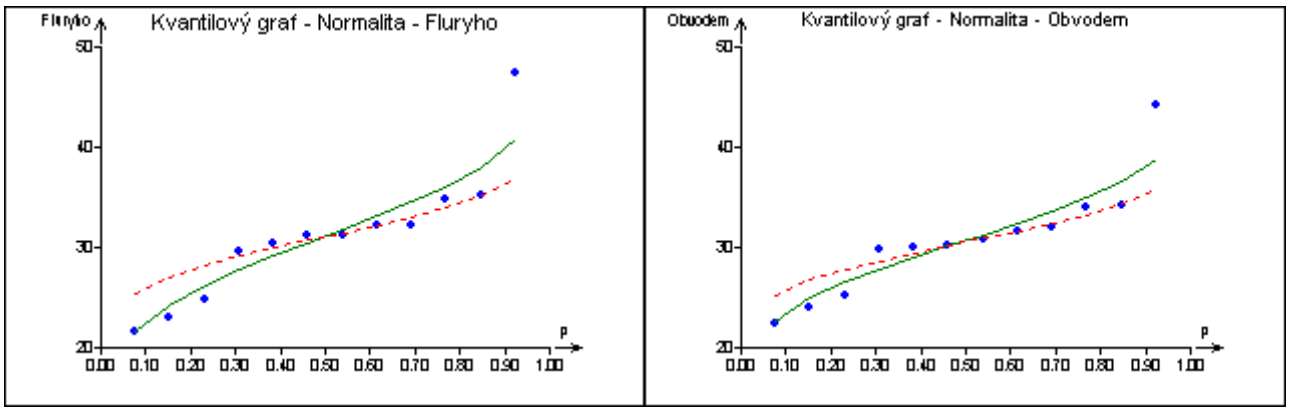
Obr. 11 Diagram rozptýlení



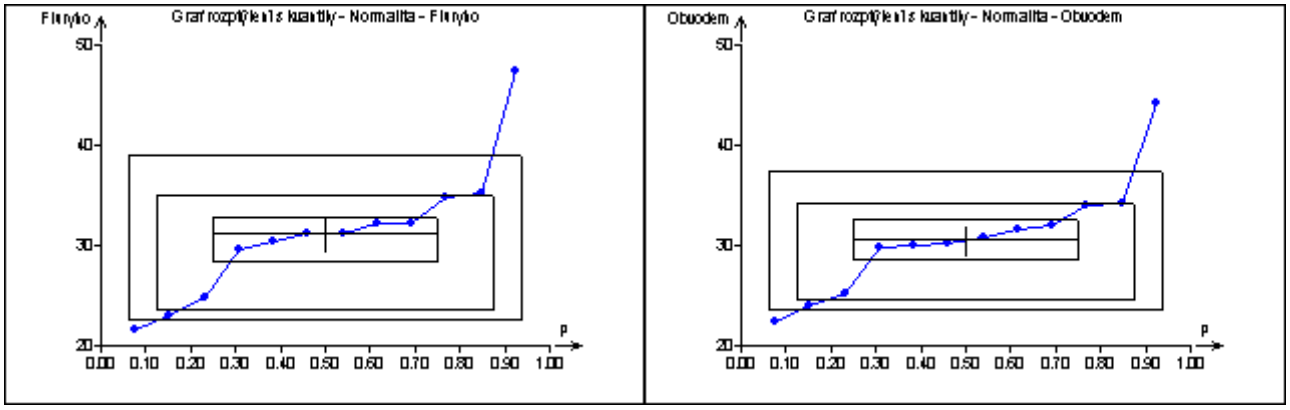
Obr. 12 Odhad hustoty



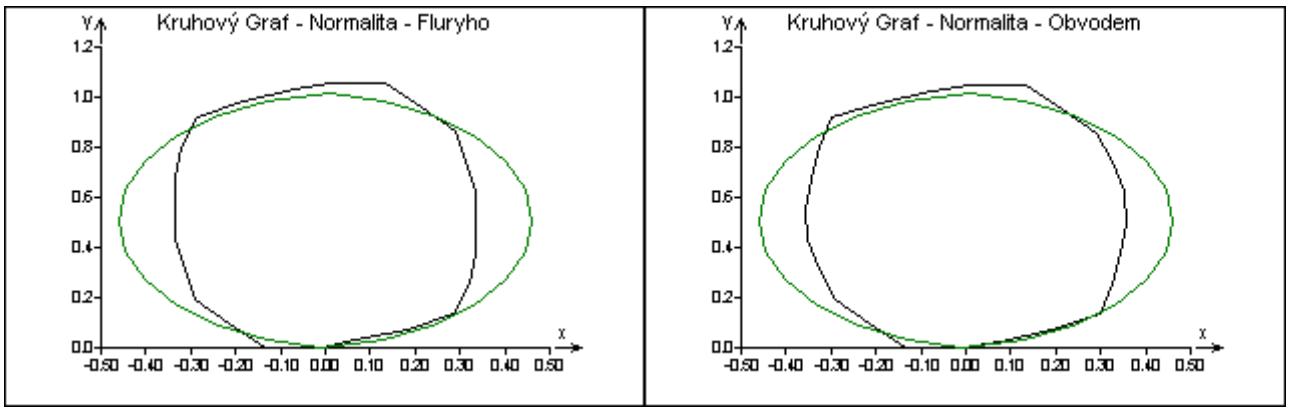
Obr. 13 Krabicový graf



Obr. 14 Kvantilový graf



Obr. 15 Graf rozptýlení s kvantily



Obr. 16 Kruhový graf

Klasické parametry :

Název sloupce :	Obvodem	Fluryho
Průměr :	30,73	31,13
Spodní mez :	27,11	26,85
Horní mez :	34,35	35,42
Rozptyl :	32,46	45,45
Směr. odchylka :	5,70	6,74
Šikmost	0,778	0,880
Odchylka od 0 :	Nevýznamná	Nevýznamná
Špičatost :	3,96	4,15
Odchylka od 3 :	Nevýznamná	Nevýznamná
Polosuma	33,30	34,55
Modus :	30,24	31,04

Robustní parametry :

Název sloupce :	Obvodem	Fluryho
Medián :	30,55	31,10
IS spodní :	18,20	16,56
IS horní :	42,90	45,64
Mediánová směr. odchylka :	5,61	6,61
Mediánový rozptyl :	31,50	43,66

Znaménkový test :

Závěr :	Data jsou nezávislá	Data jsou nezávislá

Test normality :

Název sloupce :	Obvodem	Fluryho
Průměr :	30,73	31,13
Rozptyl :	32,46	45,45
Šikmost	0,78	0,88
Špičatost :	3,96	4,15
Normalita :	Přijata	Přijata
Vypočtený T :	2,26	2,74
Teoretický T :	5,99	5,99
Pravděpodobnost :	0,32	0,25

Vybočující body :

Název sloupce :	Obvodem	Fluryho
Homogenita :	Přijata	Zamítnuta
Počet vybočujících bodů :	0	1

Porovnání dvou výběrů

Hladina významnosti :	0,05

Test shody rozptylů

Poměr rozptylů :	1,40
Kritická hodnota :	2,82
Závěr :	Rozptyly jsou SHODNÉ
Pravděpodobnost :	0,29

Test shody průměrů pro SHODNÉ rozptyly

t-statistika :	
Počet stupňů volnosti :	22,00
Kritická hodnota :	2,07
Závěr :	Průměry jsou SHODNÉ
Pravděpodobnost :	0,88

Závěr: Pro obě metody stanovení střední hodnoty průměru smrku měřením na kmeni 1.3 nad zemí stromů, data mají normální rozdělení, co je potvrzené testem normality a EDA. Při porovnání středních hodnot analyzovaných skupin test shody rozptylů potvrdil jeho shodnost. Při posuzování středních hodnot byl vykonaný test shody průměrů, který deklaroval shodnost naměřených hodnot obou metod.

(c) Párový test:

Zadání: Dendrometrem – přístrojem, který je páskem obvedeným kolem stromu a instalovaným nastálo přes celu vegetační sezonu, měří se obvod kmenu 1.3 m nad zemí a přepočítá se na průměr. Pod vlivem vlhké a suchého období (když prší nebo sucho), a času (narůstá nova biomasa) obvod se zvětšuje nebo se zmenšuje. Zjistíte vliv stress suchem na porost. Měřeni jsou do a po suché období. Zda rozdíl hodnot po stress suchem je statistický významný?

Data: Obvody stromů buku lesního do a po suchém období [cm]:

Do:

32,8	27,1	54,4	44,7	28,7	30,4	19,8	34,0	45,6	72,3	52,6	20,8

Po:

32,2	27,0	53,3	44,2	28,6	30,3	19,8	33,8	45,0	71,1	51,9	20,8

Program: ADSTAT

Řešení: **porovnání dvou výběrů, párový test**

Průměrná diference :	0,43
Rozptyl :	0,181
Počet stupňů volnosti :	11
Kritická hodnota :	2,2010
t-statistika :	8,299
Vypočtená hladina významnosti :	0,000
Závěr :	Průměry se považují za rozdílné, H_0 zamítnuta

Závěr: Párový test zamítl hypotézu o shodě obvodů stromů buku lesního do a po stress suchem období.