UNIVERZITA PARDUBICE
Fakulta chemicko-technologická
Katedra analytické chemie
Nám. Čs. Legií 565, 532 10 Pardubice

10. licenční studium chemometrie

STATISTICKÉ ZPRACOVÁNÍ DAT

Semestrální práce

STATISTICKÁ ANALÝZA JEDNOROZMĚRNÝCH DAT

2004/2005

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

Vyučující:
Prof. RNDr. Milan Meloun, DrSc.

Autor práce:
Ing. Zdeňka Dluhošová

OBSAH

ÚLOHA 1.: STATISTICKÁ ANALÝZA VELKÝCH VÝBĚRŮ

Zadání

V rámci projektu Monitorování jakosti pitné vody bylo v laboratoři analýz vod vyšetřeno 90 vzorků pitné vody na obsah hliníku. Vzorky vody byly odebrány na náhodně vybraných místech vodovodního řádu. Proveďte analýzu náhodného výběru v pořadí:

1. Průzkumová analýza dat (EDA)
2. Ověření předpokladů o výběru
3. Transformace dat
4. Statistická analýza jednorozměrných dat

Rozeberte a vysvětlete jednotlivé diagnostiky a učiňte závěry o výběru a o nejlepších odhadech polohy, rozptýlení a tvaru.

Data

Obsah hliníku v pitné vodě [mg.l^{-1}]:

0.424	0.080	0.100	0.158	0.641	0.033	0.256	0.563	0.561	0.126
0.053	0.187	0.096	0.037	0.070	0.029	0.016	0.201	0.196	0.272
0.061	0.252	0.196	0.204	0.162	0.129	0.078	0.015	0.305	0.201
0.072	0.178	0.147	0.187	0.147	0.058	0.131	0.575	0.600	0.055
0.558	0.063	0.442	0.049	0.041	0.044	0.380	0.630	0.179	0.179
0.304	0.034	0.023	0.199	0.078	0.095	0.141	0.090	0.217	0.045
0.044	0.125	0.198	0.006	0.325	0.149	0.176	0.293	0.275	0.690
0.813	0.859	0.209	0.437	0.280	1.241	1.368	1.833	0.859	0.181
0.091	0.090	0.212	0.139	0.163	0.191	0.204	0.222	0.836	0.835

Užitý program

ADSTAT: Jednorozměrná data:
Exploratorní analýza
Základní předpoklady
Mocninná transformace
Analýza 1 výběru

QC-Expert: Základní statistika (pro grafické výstupy)

Řešení

I. Exploratorní analýza

Počet dat:90
Hladina významnosti alfa: 0,050

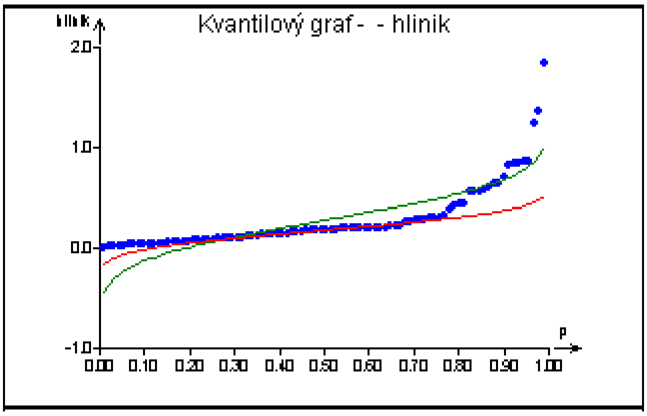
KLASICKÉ ODHADY PARAMETRŮ

Medián	: 1.8400E-01	Průměr	: 2.8876E-01
Rozptyl	: 1.2375E-01	III.centr.moment	: 1.1430E-01
IV.centr.moment	: 1.6284E-01	Šikmost	: 2.6404E+00
Špičatost	: 1.0753E+01	Směrodatná odchylka	: 3.5178E-01

Na základě hodnot šikmosti a špičatosti lze usuzovat na asymetrické, silně zešikmené a silně špičaté rozdělení dat.

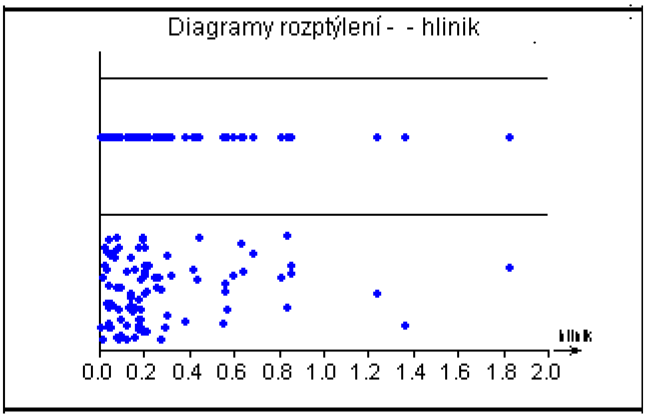
GRAFICKÉ VÝSTUPY (QC-Expert)

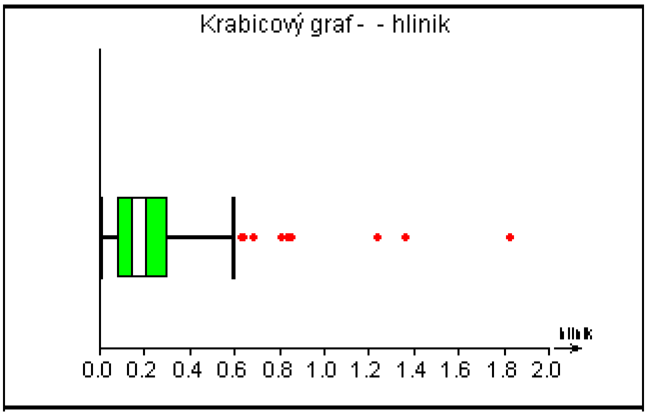
Kvantilový graf:



Je patrný velký rozdíl mezi symetrickým Gaussovým a empirickým rozdělením. Tvar křivky je charakteristický pro asymetrické rozdělení silně zešikmené k vyšším hodnotám.

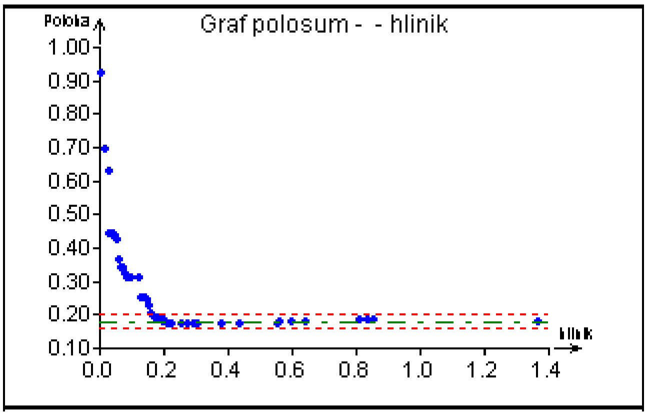
Diagram rozptýlení a krabicový graf:





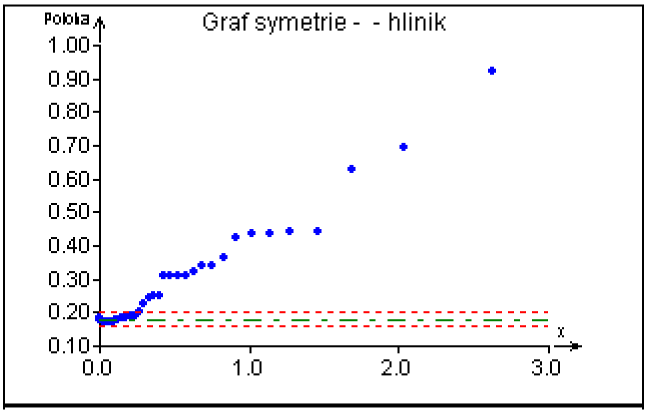
Velké zhuštění dat v oblasti nízkých hodnot, úzká krabice rozdělena na dvě části mediánem, krátké asymetrické fousy, indikovány odlehlé body, které nemůžeme vypustit – jedná se o stopovou analýzu. Graf indikuje asymetrické rozdělení dat.

Graf polosum:



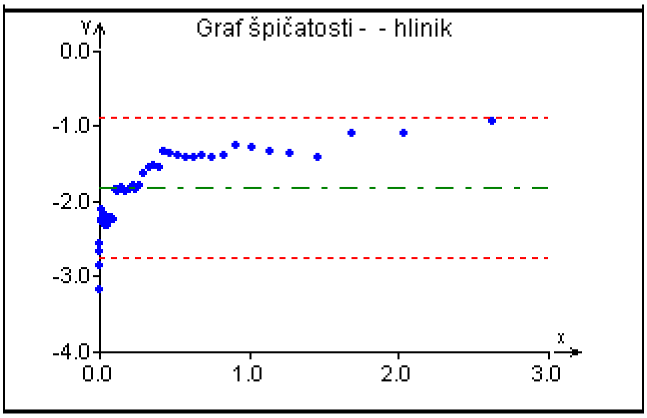
Velká část bodů se nachází mimo rovnoběžku s osou x, graf indikuje asymetrické rozdělení dat.

***Graf symetrie*:**



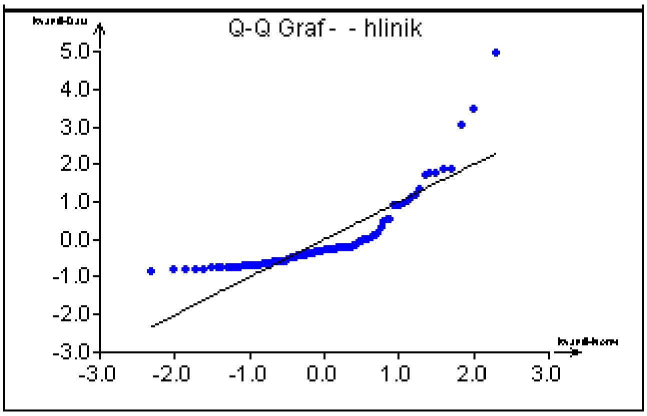
Graf indikuje velkou část bodů mimo rovnoběžku s osou x, jedná se o asymetrické rozdělení dat.

Graf špičatosti:



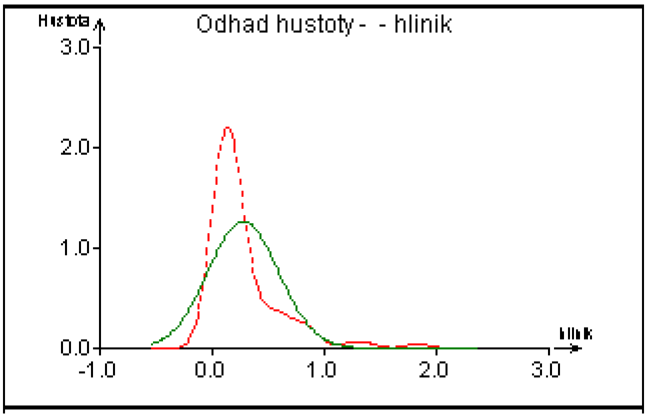
Většina bodů neleží na rovnoběžce s osou x, jedná se o asymetrické rozdělení dat.

Q-Q graf:



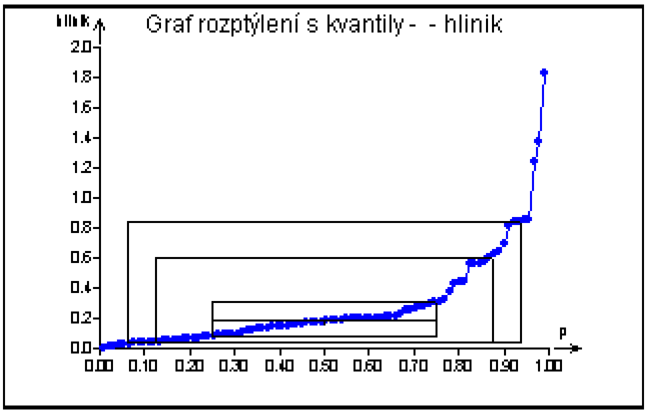
Téměř všechny body leží mimo přímku, jde o asymetrické rozdělení dat.

Graf hustoty:



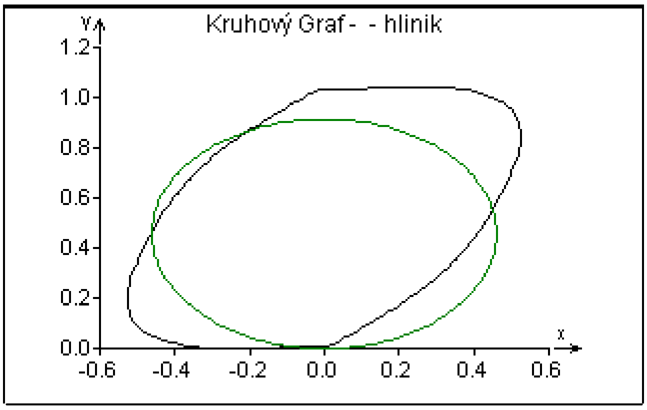
Empirická křivka je značně odlišná od křivky symetrického Gaussova rozdělení dat. Rozdělení dat je asymetrické, velmi ostré a špičaté, silně zešikmené k vyšším hodnotám.

Graf rozptýlení s kvantily:



Asymetrie kvantilových obdélníků dokazuje silně asymetrické rozdělení dat. Indikovány odlehlé body.

Kruhový graf:



Elipsa dokazuje silně asymetrické rozdělení dat zešikmené k vyšším hodnotám.

Závěr exploratorní analýzy
Na základě hodnot šikmosti a špičatosti a posouzení exploratorních (průzkumových) grafů lze usuzovat, že se jedná o silně asymetrické rozdělení dat, silně špičaté, zešikmené k vyšším hodnotám.
Byly nalezeny podezřelé body v horní části, které by mohly být chápány v symetrickém rozdělení dat jako odlehlé. Jelikož se jedná o asymetrické rozdělení, nemůžeme tato data vyloučit. Vyloučili bychom totiž data, která do výběru naopak patří, připravili bychom se tak o podstatnou informaci o tvaru rozdělení dat.
Použití aritmetického průměru jako odhadu střední hodnoty by bylo nesprávné, správnějším odhadem bude použití mediánu nebo retransformovaného průměru po provedení mocninné transformace dat.

II. Základní předpoklady

KLASICKÉ ODHADY PARAMETRŮ
Šikmost : 2.6404E+00
Špičatost : 1.0753E+01

TEST NORMALITY
Tabulkový kvantil Chi^2(1-alfa, 2) : 5.9915E+00
Chi^2-statistika : 3.8222E+02

Závěr: Předpoklad normality zamítnut.

TEST NEZÁVISLOSTI
Tabulkový kvantil t(1-alfa/2, n+1) : 1.9864E+00
Test autokorelace : 1.0149E+00

Závěr: Předpoklad nezávislosti přijat.

DETEKCE ODLEHLÝCH BODŮ
Počet odlehlých bodů: 8
Body č. 71, 72, 76, 77, 78, 79, 89, 90

Závěr základních předpokladů
Hodnota šikmosti g_1 = 2.64 poukazuje na silně zešikmené rozdělení dat (pro Gaussovo normální rozdělení se hodnota šikmosti pohybuje v intervalu <-0.3, 0.3>).
Hodnota špičatost g_2 = 10,7 poukazuje na silně špičaté rozdělení dat (pro Gaussovo normální rozdělení se hodnota špičatosti rovná 3).
Testem normality bylo potvrzeno, že se nejedná o normální rozdělení dat.
U analyzovaného výběru dat byla nezávislost prvků ve výběru prokázána.
Bylo indikováno 8 odlehlých bodů, ale tyto nelze vzhledem ke stopové analýze dat vyloučit.

III. Mocninná transformace

Prostá mocninná transformace

Optimální hodnoty mocniny pro vybraná kritéria:

Optimální mocnina: 1.3333E-01 pro šikmost	: 2.4343E-01
Optimální mocnina: 1.3333E-01 pro špičatost	: 3.0627E+00
Optimální mocnina: 1.3333E-01 pro asymetrii	: 5.7078E-04
Optimální mocnina: 6.6667E-01 pro asymetrii, rob.	: 1.2838E-03
Optimální mocnina: 1.3333E-01 pro Hinkley-asymetrii	: 1.3377E-02

Zvolená mocnina **: 0.13**

Průměr	: 7.9528E-01
Rozptyl	: 1.3259E-02
Směr. odchylka	: 1.1515E-01
Šikmost	: 2.4343E-01
Špičatost	: 3.0627E+00
Opravený průměr	: 1.7944E-01

Box-Coxova transformace

Optimální hodnoty mocniny pro vybraná kritéria:

Optimální mocnina: 1.3333E-01 pro šikmost	: 2.4343E-01
Optimální mocnina: 1.3333E-01 pro špičatost	: 3.0627E+00
Optimální mocnina: 1.3333E-01 pro asymetrii	: 5.7078E-04
Optimální mocnina: 6.6667E-01 pro asymetrii, rob.	: 1.2838E-03
Optimální mocnina: 4.0000E+00 pro Hinkley-asymetrii	: 5.4697E-02
Optimální mocnina: 1.3333E-01 pro věrohodnost	: 1.5331E+02

Zvolená mocnina **: 0.13**

Průměr	: -1.5354E+00
Rozptyl	: 7.4570E-01
Směr. odchylka	: 8.6359E-01
Šikmost	: 2.4343E-01
Špičatost	: 3.0627E+00
Opravený průměr	: 1.7944E-01

Graf věrohodnosti: Interval $<\lambda_{D,} \lambda_H>$ neobsahuje číslo +1.

Závěr mocninné transformace

Prostá mocninná transformace slouží k zesymetričtění rozdělení dat, stejně tak Box-Coxova transformace se užívá pro přiblížení rozdělení výběru k rozdělení normálnímu. Graf věrohodnosti potvrzuje vhodnost a smysl provedené transformace.

IV. Analýza 1 výběru

PARAMETRY TVARU
Šikmost : 2.6403E+00
Špičatost : 1.0753E+01

KLASICKÉ ODHADY PARAMETRŮ
Poznámka: Pouze pro ilustraci. Nelze použít pro lognormální rozdělení dat!
Průměr : 2.8876E-01
Směrodatná odchylka průměru : 3.5178E-01
Rozptyl průměru : 1.2375E-01
95 % spolehlivost:
Spodní mez : 2.1508E-01
Horní mez : 3.6243E-01

ROBUSTNÍ ODHADY PARAMETRŮ
Medián : 1.8000E-01
Směrodatná odchylka mediánu : 1.8953E-01
Rozptyl mediánu : 3.5923E-02
95 % spolehlivost:
Spodní mez : 1.5064E-01
Horní mez : 2.0936E-01

Závěr úlohy 1.

Výše uvedená data zjištěná vyšetřením vzorků pitné vody na obsah hliníku nepocházejí z rozdělení normálního. Data pochází z asymetrického rozdělení. Na základě porovnání rozdělení a nejvyššího korelačního koeficientu se jedná o lognormální rozdělení dat (korelační koeficient 9.8369E-01). Teorii lognormálního rozdělení dat lze přijmout i k jejich povaze – jedná se o stopovou analýzu kovů. Body indikovány jako odlehlé musíme plně akceptovat vzhledem k povaze rozdělení.
Pro zjištěné lognormální rozdělení dat je nejlepším odhadem polohy medián. V žádném případě nelze použít aritmetický průměr. Pro nejlepší odhad rozptýlení zvolíme rozptyl, popř. směrodatnou odchylku mediánu. Pro nejlepší odhad tvaru použijeme parametry šikmost a špičatost.
V našem případě lze jako odhad polohy použít i retransformovaný průměr po prosté mocninné transformaci nebo Box-Coxově transformaci, které vedou k výraznému přiblížení transformovaného rozdělení k normálnímu.

Medián : 0.180 mg.l^{-1}
Rozptyl mediánu : 0.036 mg^{2}l^{-2}
Směrodatná odchylka mediánu : 0.190 mg.l^{-1}
95% spolehlivost:
Spodní konfidenční mez : 0.151 mg.l^{-1}
Horní konfidenční mez : 0.209 mg.l^{-1}

Šikmost : 2.64
Špičatost : 10.75

Retransformovaný průměr : 0.179 mg.l^{-1}

<u>Úloha 2.</u>: Statistická analýza malých výběrů dle Horna

Zadání

Opakovanou analýzou referenčního materiálu byl stanovován obsah benzo(a)pyrenu. Aplikujte Hornovu metodu pivotů k určení parametrů polohy a rozptýlení a výsledky porovnejte s klasickými a robustními odhady polohy a rozptýlení pomocí software, např. ADSTAT.

Data

Obsah benzo(a)pyrenu v referenčním materiálu [ng/g]:

470	456	483	478	449	451	465	460	472	466

Užitý program

I. EXCELL
II. ADSTAT: Jednorozměrná data:
Exploratorní analýza
Základní předpoklady
Analýza 1 výběru

Řešení

I. Hornův postup

1. Pořádková statistika

i	1	2	3	4	5	6	7	8	9	10
$\mathbf{x_{(i)}}$	449	451	456	460	465	466	470	472	478	483

2. Hloubka pivotu H

H = (int ((n+1)/2)+1)/2 = (int ((10+1)/2)+1)/2 = 3

3. Dolní pivot x_D, horní pivot x_H

$x_D = x_{(H)} = x_{(3)} = 456$

$x_H = x_{(n+1-H)} = x_{(10+1-3)} = x_{(8)} = 472$

4. Pivotová polosuma P_L (odhad parametru polohy)

$P_L = (x_D + x_H)/2 = (456 + 472)/2 = 464$

5. Pivotové rozpětí R_L (odhad parametru rozptýlení)

$R_L = x_H - x_D = 472 - 456 = 16$

6. 95%interval spolehlivosti střední hodnoty

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

$464 - 16 \, . \, 0.668 \leq \mu \leq 464 + 16 \, . \, 0.668$

$453 \leq \mu \leq 475$

II. ADSTAT

Exploratorní analýza
Na základě hodnot šikmosti ($g_1 = 0.067$)a špičatosti ($g_2 = 1.95$) a posouzení exploratorních (průzkumových) grafů lze usuzovat, že se jedná o symetrické rozdělení dat. Dle hodnoty špičatosti lze předpokládat rovnoměrné rozdělení dat.

Porovnání rozdělení
Korelační koeficient rovnoměrného rozdělení dat = 9.9414E-01
Korelační koeficient normálního rozdělení dat = 9.9300E-01.

Základní předpoklady
Dle základních předpokladů byl předpoklad normality a nezávislosti přijat, soubor neobsahuje odlehlé hodnoty.

Analýza 1 výběru:

PARAMETRY TVARU
Šikmost : 0.6729E-02
Špičatost : 1.9537E+00

KLASICKÉ ODHADY PARAMETRŮ (za předpokladu normality)
Průměr : 4.6500E+02
Směr. odchylka : 1.1185E+01
Rozptyl : 1.2511E+02
95% spolehlivost:
Spodní mez : 4.5700E+02
Horní mez : 4.7300E+02

ROBUSTNÍ ODHADY PARAMETRŮ (pro neplatnost normality)
Medián : 4.6550E+02
Směr. odchylka : 1.4865E+01
Rozptyl : 2.2098E+02
95% spolehlivost:
Spodní mez : 4.5418E+02
Horní mez : 4.7682E+02

III. Porovnání výsledků dle Horna a výsledků ADSTAT

PARAMETRY	**Polohy**	**Rozptýlení**	**95% interval spolehlivosti střední hodnoty**	
Hornův postup	P_L = 464.0	R_L = 16.0	L_D = 453.0	L_H = 475.0
klasický odhad	x = 465.0	s = 11.2	L_D = 457.0	L_H = 473.0
robustní odhad	$x_{0.5}$ = 466.5	s = 14.8	L_D = 454.2	L_H = 476.8

Závěr úlohy 2.

U malých výběrů jsou závěry vždy zatíženy značnou mírou nejistoty. Malých výběrů užíváme jen tam, kde obvykle z ekonomických nebo časových důvodů není možné získat větší počet dat.
Vzhledem k počtu analýz referenčního materiálu (n = 10) 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 464 ng/g, bodový odhad rozptýlení 16 ng/g a intervalový odhad polohy je $453 \leq \mu \leq 475$.

Úloha 3.: Statistické testování

Část (a): Test správnosti

Zadání
Výrobce standardního roztoku deklaruje koncentraci fluoridů 0.50 mg/l. Jsou naměřené výsledky správné?

Data
Koncentrace fluoridů ve standardním roztoku [$mg.l^{-1}$]:

0.48	0.50	0.51	0.51	0.49	0.52	0.46	0.48	0.53	0.51

Užitý program
ADSTAT: Jednorozměrná data:
Exploratorní analýza
Základní předpoklady
Analýza 1 výběru

Řešení
Z exploratorní analýzy dat byla zjištěna mírná asymetrie, posun k nižším hodnotám, vizuálně indikován 1 odlehlý bod (9). Dle základních předpokladů byl předpoklad normality a nezávislosti přijat, soubor neobsahuje odlehlé hodnoty.

Analýza 1 výběru
Počet dat: 10
Hladina významnosti: 0.050

PARAMETRY TVARU
Šikmost : -3.7140E-01
Špičatost : 2.2373E+00

KLASICKÉ ODHADY PARAMETRŮ (za předpokladu normality)
Průměr : 4.9900E-01
Směr. odchylka : 2.1318E-02
Rozptyl : 4.5444E-04
95% spolehlivost:
Spodní mez : 4.8375E-01
Horní mez : 5.1425E-01

ROBUSTNÍ ODHADY PARAMETRŮ (pro neplatnost normality)
Medián : 5.0500E-01
Směr. odchylka : 2.7873E-02
Rozptyl : 7.7688E-04
95% spolehlivost:
Spodní mez : 4.8150E-01
Horní mez : 5.2850E-01

OSTATNÍ ODHADY POLOHY
Odhad modu : 5.1000E-01
Odhad polosumy : 4.9500E-01

Závěr úlohy 3(a).

Pro 95% statistickou jistotu byly nalezeny následující intervalové odhady - pro aritmetický průměr x je interval $0.48 \le \mu \le 0.51$, pro medián $x_{0.5}$ je interval $0.48 \le \mu \le 0.53$. Z uvedených intervalových odhadů vyplývá, že výrobcem deklarovaná koncentrace fluoridů 0.50 $mg.l^{-1}$ leží v 95% intervalu spolehlivosti střední hodnoty a naměřené hodnoty jsou správné.

Část (b): Test shodnosti

Zadání

Pro stanovení dusitanů v pitné vodě byly použity dvě metody - spektrofotometrická metoda a metoda iontové chromatografie. Stanovení byla prováděna na kontrolním vzorku vody s obsahem dusitanů 0.120 $mg.l^{-1}$. Vedou obě metody ke stejným výsledkům?

Data

Koncentrace dusitanů v kontrolním vzorku [$mg.l^{-1}$]:
data (a) – spektrofotometrická metoda
data (b) – metoda iontové chromatografie

Data (a)

0.116	0.118	0.136	0.121	0.133	0.144	0.125	0.127	0.132	0.124	0.131	0.106

Data (b)

0.124	0.118	0.123	0.117	0.130	0.119	0.127	0.125	0.114	0.126	0.119	0.123

Užitý program

ADSTAT: Jednorozměrná data:
Základní předpoklady
Porovnání 2 výběrů

Řešení

Z ověření základních předpokladů pro jednotlivé výběry vyplývá, že data v obou výběrech jsou nezávislá, test normality prokázal Gaussovo rozdělení, nejsou detekovány odlehlé hodnoty.

Porovnání 2 výběrů

KLASICKÉ ODHADY PARAMETRŮ

Parametr	**Výběr 1**	**Výběr 2**	**Celkově**
Velikost výběru	12	12	24
Průměr	1.2617E-01	1.2208E-01	1.2413E-01
Rozptyl	1.0361E-04	2.2083E-05	6.0112E-05
Šikmost	-2.3481E-01	-7.1127E-02	-2.5593E-01
Špičatost	2.7371E+00	2.0805E+00	3.8481E+00

TEST HOMOGENITY ROZPTYLU (hypotéza H0:s1^2 = s2^2):

Fischer-Snedocorův F-test

Počet stupňů volnosti Df1	11
Df2	11
Tabulkový kvantil F(1-alfa/2, Df1, Df2)	3.4737E+00
Experimentální F-statistika	4.6916E+00

Závěr: Rozptyly se považují za rozdílné, H0 zamítnuta.
Vypočtená hladina významnosti: 0.008.

TEST SHODY PRŮMĚRU (hypotéza H0: $\mu_1 = \mu_2$)

Studentův t-test (pro různé rozptyly)

Počet stupňů volnosti Df1	17
Tabulkový kvantil t (1-alfa/2, DF1)	2.1098E+00
Experimentální t-statistika	1.2617E+00

Závěr: Průměry se považují za shodné, H0 přijata.
Vypočtená hladina významnosti: 0.224.

Závěr úlohy 3(b).

Na hladině významnosti $\alpha = 0.05$ potvrzuje oboustranný klasický test shodu středních hodnot obou výběrů i při významné odlišnosti obou rozptylů. Obě metody stanovení dusitanů poskytují stejné výsledky, i když každá s jinou variabilitou.

Část (c): Párový test

Zadání

Při zkoušce spolehlivosti nového průtokoměru bylo provedeno 20 kontrolních měření při porovnání s druhým referenčním průtokoměrem. Párovým testem určete, zda nový průtokoměr poskytuje stejné výsledky jako průtokoměr referenční?

Data

Naměřený průtok [$l.min^{-1}$]:
data (a) - nový průtokoměr, data (b) – referenční průtokoměr

Data (a)

12.1	8.5	23.2	35.4	16.7	19.9	25.5	12.1	18.3	14.6
34.0	26.7	39.0	17.2	26.4	18.5	17.3	24.0	19.1	30.0

Data (b)

12.4	7.9	23.7	35.5	16.0	20.3	25.1	12.5	17.9	14.4
34.3	26.8	38.5	16.8	26.1	18.9	17.6	23.6	18.7	30.2

Užitý program

ADSTAT: Jednorozměrná data:
Porovnání 2 výběrů

Řešení

Porovnání 2 výběrů

TEST SHODY PRŮMĚRU (hypotéza H0: $\mu_1 = \mu_2$)
t-test párový

Průměrný rozdíl	6.5000E-02
Rozptyl	9.6884E-02
Počet stupňů volnosti Df1	19
Tabulkový kvantil t(1-alfa/2, Df1)	2.0930E+00
Experimentální t-statistika	3.0004E+00

Závěr: Průměry se považují za rozdílné, H0 zamítnuta.
Vypočtená hladina významnosti: 0.007.

Závěr úlohy 3(c).

Na hladině významnosti $\alpha = 0.05$ párový t-test zamítá hypotézu o shodě průměrů. Nový průtokoměr neposkytuje stejné výsledky jako průtokoměr referenční.

Použitá literatura

1. Meloun M., Militký J.: Kompendium statistického zpracování dat, Academia, Praha, 2002
2. Kupka, K.: Statistické řízení jakosti, TriloByte, Pardubice, 1997