Univerzita Pardubice

Fakulta chemicko – technologická

Katedra analytické chemie

Licenční studium chemometrie

Statistické zpracování dat

Statistická analýza jednorozměrných dat

Zdravotní ústav se sídlem v Ostravě
Odbor hygienických laboratoří Karviná

V Karviné dne 10.1.2005 **Ing. Miluše Galuszková**

Předmět:

1.3.STATISTICKÁ ANALÝZA JEDNOROZMĚRNÝCH DAT

Přednášející: ***Prof.RNDr. Milan Meloun, DrSc.***

Obsah

Úloha 1 Statistická analýza velkých výběrů

Na větší výběr dat z pracoviště aplikujte obecný postup analýzy náhodného výběru v pořadí :
1. Průzkumová analýza spojitých dat (EDA)
2. Ověření předpokladů o výběru
3. Transformace dat
4. Statistická analýza jednorozměrných dat (CDA)

Zadání

V laboratoři byl připraven homogenizací vzorek prašného spadu (10g). Po mikrovlnné mineralizaci byl metodou AAS stanoven obsah kadmia v mg/kg. Analýza byla prováděna v rámci analýz reálných vzorků v období r.1998 -2002 .
Splňují data tohoto výběru základní předpoklady na výběr?
Podle závěrů průzkumové analýzy navrhněte vhodný typ rozdělení. Proveďte analýzu výběru.

Data:

Obsah kadmia v prašném spadu v mg/kg

1.58	1.55	1.60	1.60	1.65
1.67	1.45	1.55	1.75	1.80
1.57	1.50	1.45	1.60	1.70
1.62	1.50	1.75	1.50	1.50
1.68	1.65	1.55	1.50	1.55
1.65	1.50	1.65	1.60	1.70
1.65	1.70	1.70	1.70	1.55
1.75	1.45	1.80	1.55	1.65
1.60	1.50	1.70	1.70	1.50
1.55	1.60	1.75	1.65	1.55
1.50	1.60	1.70	1.55	1.44
1.65	1.55	1.45	1.60	1.58

1. Průzkumová analýza spojitých dat (EDA)

Program: **QC – Expert**

Modul: **Základní statistika**

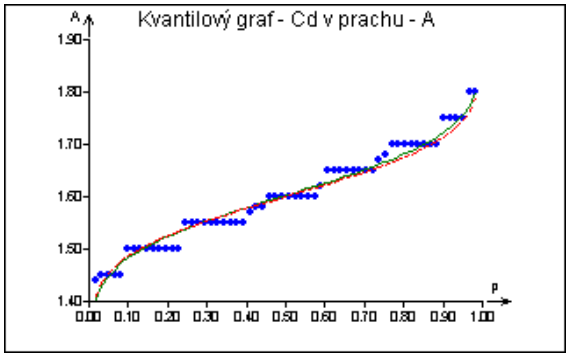
Řešení: **Diagnostické grafy**

Grafy:

Pomocí grafů indikujeme statistické zvláštnosti dat:

- stupeň symetrie rozdělení výběru
- stupeň špičatosti rozdělení výběru
- lokální koncentrace dat
- přítomnost odlehlých bodů

Obr.1.1 Kvantilový graf



Kvantilový graf: Symetrické rozdělení, pravděpodobně normální rozdělení, nejsou indikovány odlehlé body.

Obr.1.2 Diagram rozptýlení a rozmítnutý diagram rozptýlení

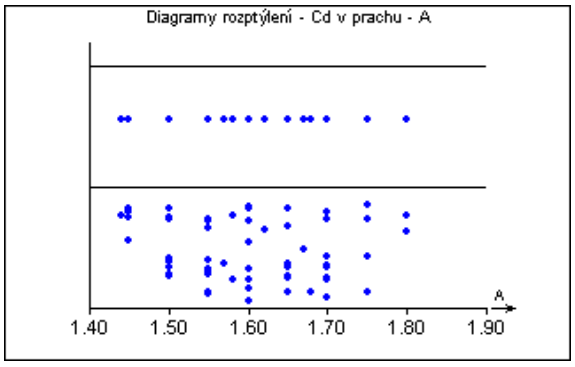
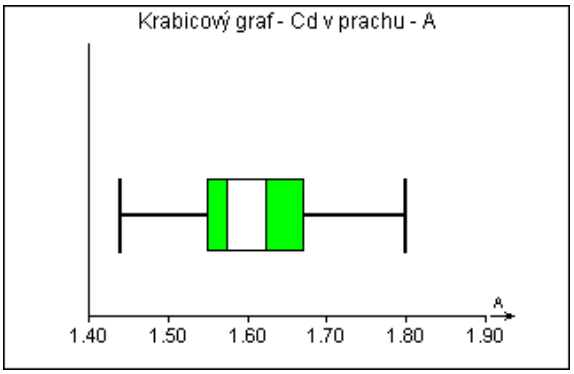


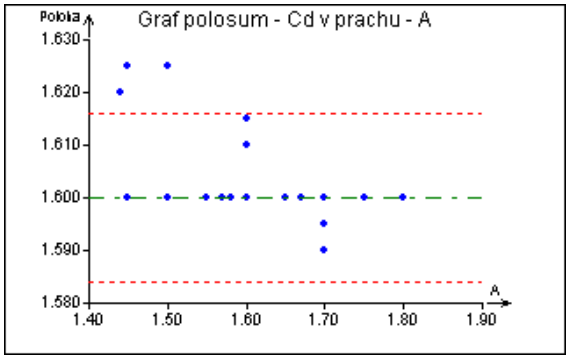
Diagram rozptýlení a rozmítnutý diagram ukazuje lokální koncentraci dat, jedná se pravděpodobně o symetrické rozdělení, 2 odlehlé body v horní části.

Obr.1.3 Krabicový graf



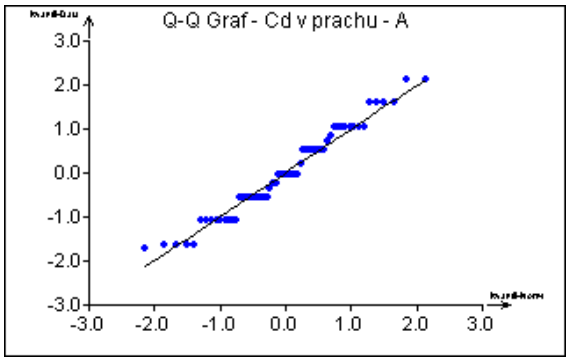
Krabicový graf neindikuje mimo vnitřní hradby odlehlé body. Kvantily nejsou symetrické.

Obr.1.4 Graf polosum



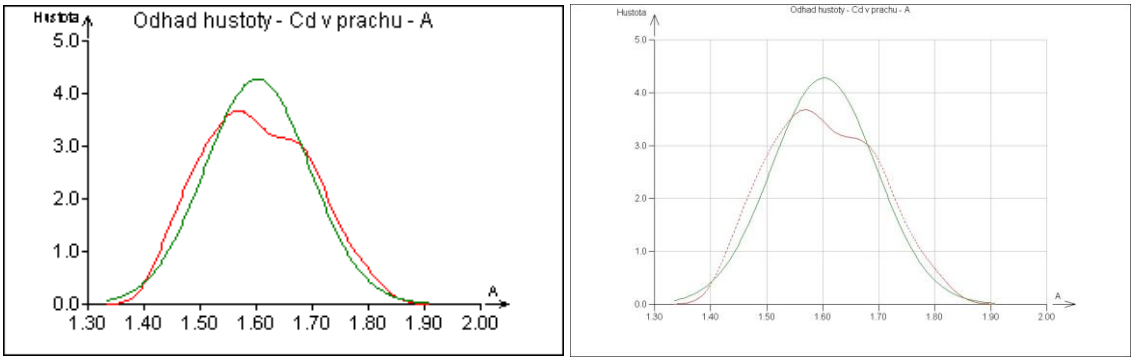
Graf polosum indikuje, že větší část bodů osciluje okolo horizontální přímky, jde o symetrické rozdělení.

Obr.1.5 Q-Q graf



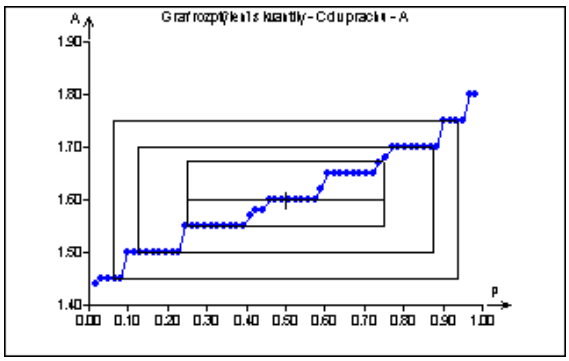
Q – Q graf znázorňuje polohů bodů na přímce, jde o symetrické rozdělení.

Obr.1.6 Graf hustoty pravděpodobnosti



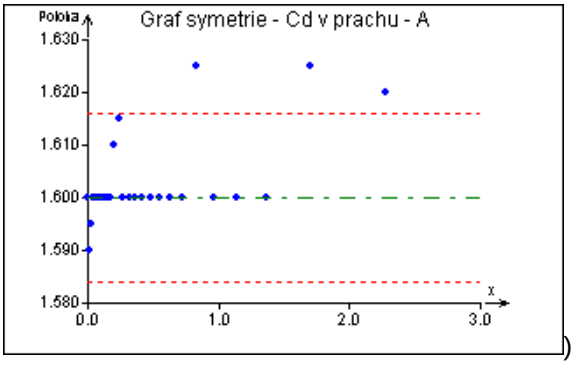
Graf jádrový odhad hustoty pravděpodobnosti ukazuje, že rozdělení není silně tvarově odlišné od normálního, je však přítomen hrb (podezření na bimodální data).
Kolmice z vrcholu Gaussovy křivky ukazuje na ose x hodnotu průměru, kolmice z vrcholu empirické křivky ukazuje na ose x hodnotu mediánu, je zřejmý mírný posun.

Obr.1.7 Graf rozptýlení s kvantily



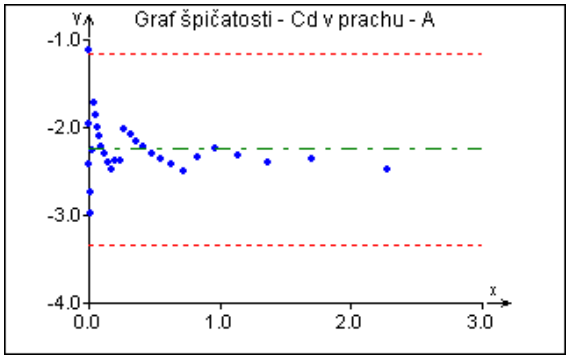
Graf rozptýlení s kvantily ukazuje na mírnou asymetrii a indikuje 3 odlehlé body.

Obr.1.8 Graf symetrie



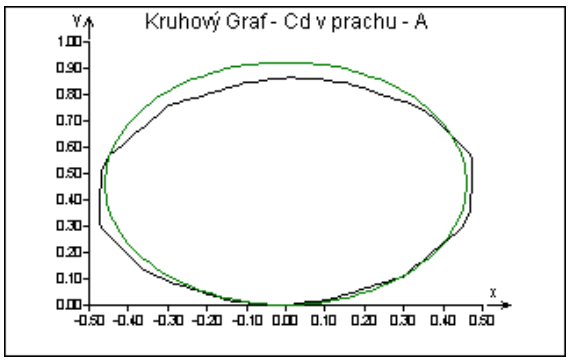
Graf symetrie indikuje velkou část bodů kolem osy x, jedná se o symetrické rozdělení.

Obr.1.9 Graf špičatosti



Graf špičatosti znázorňuje body rozložené rovnoměrně s osou x.

Obr.1.9 Kruhový graf



Kruhový graf znázorňuje empirickou křivku ve tvaru elipsy rovnoběžnou s osou x, což ukazuje na symetrické rozdělení.

Závěr:

Z diagnostických grafů vyplývá, že se jedná o symetrické rozdělení s podezřením na bimodální data. Byly indikovány 3 odlehlé hodnoty. Vzhledem k dlouhé době analýzy, data vznikala v rozmezí několika let, na základě znalosti zpracování vzorků, nedoporučuji tyto body vyřadit ze souboru.

Porovnání rozdělení:

Program: **ADSTAT 1.25**
Modul: **Jednorozměrná data**
Metoda: **Porovnání rozdělení**
Řešení: **Rozdělení s korelačním koeficientem nejblíže 1,**

Výstup:

Rozdělení	**Korelační koeficient**
Laplaceovo	0.95605
Normální	0.98515
Exponenciální	0.91851
Rovnoměrné	0.98561
Lognormální	0.83635
Gumbelovo	0.94512

Závěr porovnání rozdělení:

Podle velikosti korelačního koeficientu usuzujeme na rovnoměrné rozdělení.
Rovnoměrné rozdělení má korelační koeficient 0.98561, těsně za tímto rozdělením je s korelačním koeficientem 0,98515 rozdělení normální.

2. Ověření předpokladů o výběru

Program: **ADSTAT 1.25**
Modul: **Jednorozměrná data**
Metoda: **Základní předpoklady**
Řešení: Statistickým hodnocením ověříme, zda data jsou na sobě nezávislá, mají normální rozdělení a neobsahují-li vybočující body.

Počet dat: n = 60
Hladina významnosti: α = 0,05
Výstup

1. Klasické odhady parametrů:

Průměr	1.6023	Rozptyl	0.0087059
Směrodatná odchylka	0.093305	Šikmost	0.16806
Špičatost	2.204		

2. Test normality:

Tabulkový kvantil Chi ^2 (1-α,2)	5.9915
Chi^ 2-statistika	1.8657
Závěr: Předpoklad normality přijat.	
Vypočtená hladina významnosti:	0.39344

3. Test nezávislosti:

Tabulkový kvantil t(1-α/2, n+1)	1.9996
test autokorelace:	1.6442
Závěr: Předpoklad nezávislosti přijat.	
Vypočtená hladina významnosti:	0.052635

Předpoklad homogenity výběru:

Aritmetický průměr:	1.6023
Rozptyl:	0.0087059
Směrodatná odchylka:	0.093305

Vnitřní meze:

Spodní mez:	1.2764
Horní mez:	1.9486

4. Minimální velikost výběru:

Pro 25% relativní chybu směrodatné odchylky:	n = 6
Pro 10% relativní chybu směrodatné odchylky:	n = 31
Pro 5% relativní chybu směrodatné odchylky:	n = 121

5. Detekce odlehlých bodů:

Ve výběru nejsou odlehlé body

Závěr ověření předpokladů o výběru:

Předpoklad o normalitě výběru a nezávislosti prvků přijat. Test nenašel odlehlé body. testy jsou méně citlivé na odchylky od normality než diagnostické grafy. Vyloučením odlehlých bodů indikovaných diagnostickými grafy by mohlo dojít ke ztrátě úplné informace. Odlehlé body jsou pouze podezřelé, budou ponechány v souboru dat.
Transformace dat není nutná, lze přistoupit přímo ke statistické analýze jednorozměrných dat (CDA).

3. Transformace dat

V rámci úlohy 1.3 provedeme pokus o zlepšení rozdělení pomocí transformace.

Program: **ADSTAT 1.25**
Modul: **Jednorozměrná data**
Metoda: M**ocninná transformace**

Řešení:
(1) Analýza původních dat

(A) Klasické odhady parametrů:

Průměr:	1.6023
Rozptyl:	0.008706
Směrodatná odchylka:	0.093308
Šikmost:	0.16574
Špičatost:	2.2043

(B) Kvantilové míry:

Kvantil	P	Spodní mez	Horní mez	Polorozptyl
Medián	0.5	1.6000	-	-
Kvartil	0,25	1.5500	1.6775	0.12750

(C) míry rozptylu:

Kvantil	P	Polosuma	Šikmost	Délka konců	Norm.d.konců
Kvartil	0.25	1.6137	-0.10784	0.0000	0.0000

(2) Prostá mocninná transformace:

(A) Optimální hodnoty mocniny pro vybraná kritéria:

Optimální mocnina:	-0.66667	pro šikmost	0.0017054
Optimální mocnina:	4.000	pro špičatost	3.5122
Optimální mocnina:	0.1333	pro asymetrii	0.000072131
Optimální mocnina:	-4.00000	pro asymetrii, rob	0.11996
Optimální mocnina:	-3.8667	pro Hinkley-asym.	0.00016367

Zvolená mocnina:	-0,67
Průměr:	0.73164
Rozptyl:	0.0008017
Směrodatná odchylka:	0.028315
Šikmost:	0.0017054
Špičatost:	2.1521
Opravený průměr:	1.5979

(B) Kvantilové míry:

Kvantil	P	Spodní mez	Horní mez	Polorozptyl
Medián	0.5	0.73100	-	-
Kvartil	0.25	0.70832	0.74664	0.038326

(C) míry rozptylu:

Kvantil	P	Polosuma	Šikmost	Délka konců	Norm.d.konců
Kvartil	0.25	0.72748	0.091999	0.0000	0.0000

(3) Box - Coxova transformace:

(A) Optimální hodnoty mocniny pro vybraná kritéria:

Optimální mocnina:	-0.66667	pro šikmost	0.0017054
Optimální mocnina:	4.000	pro špičatost	3.5122
Optimální mocnina:	0.1333	pro asymetrii	0.000072131
Optimální mocnina:	-4.0000	pro asymetrii, rob	0.11996
Optimální mocnina:	-3.8667	pro Hinkley-asym.	0.000042328
Optimální mocnina:	-0.5333	pro věrohodnost.	142.56

Zvolená mocnina:	-0,67
Průměr:	0.40253
Rozptyl:	0.001839
Směrodatná odchylka:	0.042472
Šikmost:	-0.0017054
Špičatost:	2.1521
Opravený průměr:	1.5979

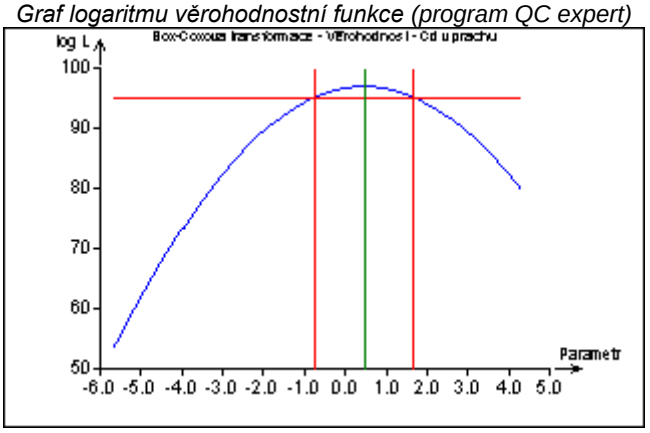
(B) Kvantilové míry:

Kvantil	P	Spodní mez	Horní mez	Polorozptyl
Medián	0.5	0.40349	-	-
Kvartil	0.25	0.38004	0.43753	0.0057490

(C) míry rozptylu:

Kvantil	P	Polosuma	Šikmost	Délka konců	Norm.d.konců
Kvartil	0.25	0.40878	-0.09199	0.0000	0.0000

Graf logaritmu věrohodnostní funkce (program QC expert)



Závěr transformace dat:

K zesymetričtění rozdělení výběru byla využita prostá mocninná transformace.
Optimální odhad koeficientu λ ze selekčního grafu dle Hinese-Hinesové λ= - 0.67.

Pro přiblížení výběru k normálnímu byla použita Box-Coxova transformace.
Z grafu logaritmu věrohodnostní funkce je patrné, že 95%ní konfidenční parametru λ je dosti široký a zahrnuje hodnotu 1, proto transformace ze statistického hlediska není přínosem.

5. Statistická analýza jednorozměrných dat (CDA)

Program: **ADSTAT 1.25**

Modul: **Jednorozměrná data**

Metoda: **Analýza 1 výběru**

Výstup:

(1) Parametry tvaru	Šikmost	0.16574		
	Špičatost	2.204		
(2) Klasické odhady parametrů				
	Průměr:	1.6023		
	Směrodatná odchylka:	0.093308		
	Rozptyl:	0.008706		
	95% spolehlivost:			
	Spodní mez:	1.5782	Horní mez:	1.6264
(3) Ostatní odhady polohy:	Odhad modu:	1.5000		
	Odhad polosumy:	1.6200		
(4) Robustní odhady parametrů:	Medián:	1.6000		
	Směr. odchylka med.:	0.11149		
	Rozptyl mediánu:	0.012430		
	Rozptyl (nepar):	0.00065077		
	Rozptyl (Marritz):	0.00038565		
	Směr.odchylka med.:	0.019638		
	95% spolehlivost:			
	Spodní mez:	1.5607	Horní mez:	1.6393
Uřezání 5% (pro P = 0,05)				
	Průměr:	1.600		
	Směrodatná odchylka:	0.09980		
	Rozptyl:	0.0099761		
	Průměr, winsor.:	1.6008		
	St.odch.winsor.	0.093873		
	Rozptyl, winsor.:	0.008812		
	95% spolehlivost:			
	Spodní mez:	1.5751	Horní mez:	1.6268
Uřezání 10% (pro P = 0,10)				
	Průměr:	1.600		
	Směrodatná odchylka:	0.094868		
	Rozptyl:	0.009000		
	Průměr, winsor.:	1.6000		
	St.odch.winsor.:	0.083964		
	Rozptyl, winsor.:	0.0.0070500		
	95% spolehlivost:			
	Spodní mez:	1.5754	Horní mez:	1.6268
Biweight (Robustní M odhady)				
	Průměr:	1.60009		
	Směrodatná odchylka:	0.096230		
	Rozptyl:	0.0092603		
	Váhy sqrt (W):	7.4285		
	95% spolehlivost:			
	Spodní mez:	1.575	Horní mez:	1.6268

Závěr:

Odhady parametrů polohy a rozptýlení byly ovlivněny odchylkami od normality v datech (vybočující hodnoty, podezření na bimodální data). Proto je výhodnější využít robustní M odhady .
Průměrný obsah kadmia (g/kg) v prašném spadu je 1.60009 se směrodatnou odchylkou 0.09623.

Úloha 2 Statistická analýza malých výběrů dle Horna

Zadání:

Vzorek moči byl po rozmražení homogenizován. V tomto vzorku byl 8x stanoven fotometrickou metodou obsah kreatininu.
Aplikujte Hornovu metodu pivotů k určení parametrů polohy a rozptýlení a výsledky porovnejte s klasickými a robustními odhady polohy a rozptýlení pomocí software ADSTAT

Data:

n = 8

x (g/l)	1.40	1.44	1.45	1.42	1.50	1.48	1.38	1.35

1. Hornův postup pivotův pro malé výběry (4<n<20)

1.1 Pořádkové statistiky

i	1	2	3	4	5	6	7	8
x_i	1.35	1.38	1.40	1.42	1.44	1.45	1.48	1.50

1.2. Hloubka pivotu

$$H = \frac{\frac{n+1}{2}+1}{2} = \text{int}(2{,}75) \approx 2$$ **H = 2**

1.3. Pivoty

Dolní pivot $x_D = x_H$ **$x_{(2)}$ = 1.38**

Horní pivot $x_H = x_{(n+1-H)}$ **$x_{(7)}$ = 1.48**

1.3. Pivotová polosuma

$$P_L = \frac{x_D + x_H}{2}$$ **P_L = 1.43**

1.5. Pivotové rozpětí

$$R_L = x_H - x_D$$ **R_L = 0.10**

1.6. 95%ní interval spolehlivosti střední hodnoty μ

$$P_L - R_L \cdot t_{L,1-\alpha/2}(n) \le \mu \le P_L + R_L \cdot t_{L,1-\alpha/2}(n)$$

1.43 – 0.10·0.564 ≤ μ ≤ 1,43 + 0.10·0.564 **1.37 ≤ μ ≤ 1.49**

1.7. Závěr

Bodový odhad míry polohy je 1.43 , míry rozptýlení 0.10 a intervalový odhad polohy je 1.37 ≤ μ ≤1.49.

2. Klasické a robustní odhady polohy a rozptýlení pomocí software ADSTAT 1.25

Modul: **Jednorozměrná data**

Metoda: **Analýza 1 výběru**

Základní analýza dat

V S T U P

Počet dat:	8
Hladina významnosti alfa :	0.050
Název výstupního souboru :	KREATIN.TXT

V Ý S T U P

(1) Parametry tvaru

Šikmost:	-0.065483
Špičatost:	1.9614

(2) Klasické odhady parametrů:

Průměr:	1.4275		
Směr.odchylka:	0.050356		
Rozptyl:	0.0025357		
95.0% spolehlivost:			
Spodní mez:	1.3854	Horní mez:	1.4696

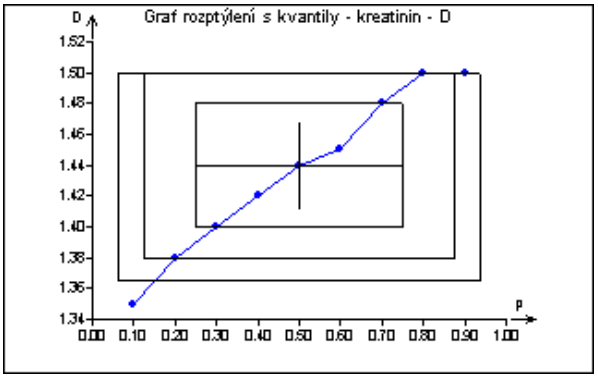
(3) Ostatní odhady polohy:

Odhad modu:	1.4100
Odhad polosumy:	1.4250

(4) Robustní odhady parametrů:

Medián:	1.4300		
Směr. odchylka mediánu:	0.074327		
Rozptyl:	0.0013448		
Směr. odchylka mediánu:	0.024982		
95.0% spolehlivost:			
Spodní mez:	1.3709	Horní mez:	1.4891

Obrázek 2.1: Kvantilový graf (program QC Expert)



3. Porovnání výsledků Hornova postupu s klasickými a robustními odhady polohy a rozptýlení

Parametry	polohy	rozptýlení	95%ní interval spolehlivosti	
Klasické odhady	$\bar{x}$ = 1.428	s = 0.050	L_D = 1.385	L_H = 1.470
Robustní odhady	$\tilde{x}_{0,5}$ =1.430	s = 0.074	L_D =1.371	L_H = 1.49
Hornův postup	P_L =1.43	R_L = 0.10	L_D = 1.37	L_H = 1.489

4. Závěr

Předpoklad normality a nezávislosti byl přijat.

Vzhledem k počtu měření a na základě porovnání s klasickými a robustními odhady polohy a rozptýlení lze přijmout závěr Hornova postupu.

Střední hodnota koncentrace (g/l) kreatininu v moči leží s 95%ní pravděpodobností v intervalu 1.37 do 1.49.

Úloha 3 Statistické testování

Část (a) Test správnosti

Zadání:

V laboratoři bylo provedeno stanovení obsahu etanolu ve vzorcích alkoholického nápoje. Výrobcem deklarovaná koncentrace byla 37.5%. Ověřte zda deklarovaná koncentrace byla správná.
Stanovení bylo provedeno 10x. Aplikujte na tato data test správnosti.

Data:

n = 10

x (%)	36.9	37.5	37.2	37.6	37.4	36.5	37.6	36.2	37.9	37.4

Vzhledem k malému počtu dat je výhodné využít Hornův postup pivotův pro malé výběry (4<n<20)

Pořádkové statistiky

i	1	2	3	4	5	6	7	8	9	10
x_i	36.2	36.5	36.9	37.2	37.4	37.4	37.5	37.6	37.6	37.9

Hloubka pivotu

$$H = \frac{\frac{n+1}{2}+1}{2} = \text{int}(3.25) \approx 3$$ **H = 3**

Pivoty

Dolní pivot $x_D = x_H$ **$x_{(3)}$ = 36.9**

Horní pivot $x_H = x_{(n+1-H)}$ **$x_{(8)}$ = 37.6**

Pivotová polosuma

$$P_L = \frac{x_D + x_H}{2}$$ **P_L = 37.25**

Pivotové rozpětí

$$R_L = x_H - x_D$$ **R_L = 0.70**

95%ní interval spolehlivosti střední hodnoty μ

$$P_L - R_L \cdot t_{L,1-\alpha/2}(n) \le \mu \le P_L + R_L \cdot t_{L,1-\alpha/2}(n)$$

37.25 – 0.70·0.668 ≤ μ ≤ 37.25 + 0.70·0.668

36.78 ≤ μ ≤ 37.72

Závěr

Interval spolehlivosti obsahuje hodnotu 37.5, proto lze konstatovat, že výrobcem deklarovaná hodnota je správná.

Část (B) Test shodnosti

Zadání

Ve vzorku odpadní vody byl stanoven arzén metodami AAS:
metoda M1: AAS-VGA (generace hydridů)
metoda M2: AAS- GTA (elektrotermická atomizace).

Aplikujte test shodnosti stanovení obsahu arzénu u obou metod na hladině významnosti α = 0.05.

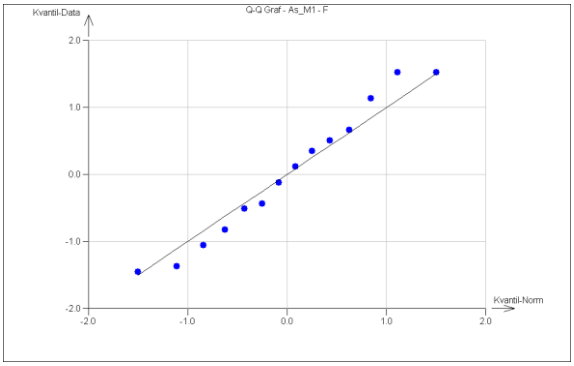
Data:

n = 14

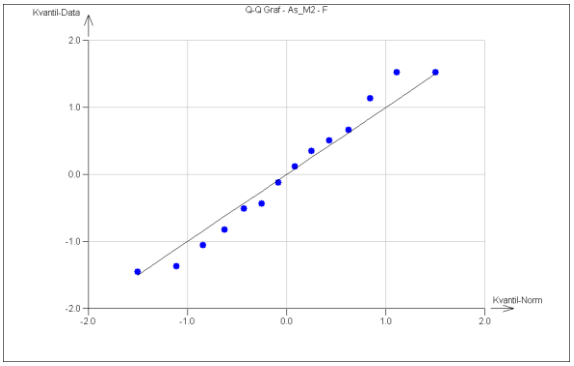
M1	34.0	35.2	35.0	34.5	36.0	35.5	37.4	35.0	33.5	35.9	33.5	35.8	34.9	36.0
M2	37.0	37.5	34.5	34.2	37.5	35.7	35.4	33.8	33.7	35.0	36.2	34.9	36.0	36.4

Z ověření základních předpokladů pro jednotlivé výběry vyplývá, že data v obou výběrech jsou nezávislá, homogenní a neobsahují odlehlé body, předpoklad normality výběru přijat.

Q-Q graf : metoda M1 (program QCExpert)



Q-Q graf : metoda M2(program QCExpert)



Q – Q graf znázorňuje polohů bodů na přímce, jde o symetrické rozdělení.

Z ověření základních předpokladů pro jednotlivé výběry vyplývá, že data v obou výběrech jsou nezávislá, homogenní a neobsahují odlehlé body, předpoklad normality výběru přijat.

Testování:

Program: **ADSTAT 1.25**
Modul: **Jednorozměrná data**
Metoda: **Porovnání dvou výběrů**

Řešení:

(1) Klasické odhady parametrů:

Parametr	M1	M2	Celkově
Velikost výběru	14	14	28
Průměr	35.157	35.557	35.357
Rozptyl	1.1473	1.6334	1.3388
Šikmost	0.1453	0.10146	0.11905
Špičatost	2.7612	1.8598	2.2235

(2) Test homogenity rozptylu (hypotéza H0: s1^2 = s2^2)

Korigovaný F-test:

Počet stupňů volnosti	Df1: 21
	Df2: 21
Tabulkový kvantil F(1-α/2, Df1, Df2)	2.4086
F-statistka	1.4238

Závěr:
Rozptyly se považují za shodné, H0 přijata.

Vypočtená hladina významnosti: 0.212

(3) Test shody průměru (hypotéza H0: μ1 = μ2):

t – test (modifikovaná šikmost)

U – star:	0.8967
B – X charakteristika:	0.0044786
B – Y charakteristika:	0.0053126
Počet stupňů volnosti	Df1: 27
Tabulkový kvantil F(1-α/2, Df1)	2.0518
t – statistika:	0.8967

Závěr:

Průměry se považují za shodné, H0 přijata.

Vypočtená hladina významnosti: 0.378

Závěr:

Na zvolené hladině významnosti α=0.05 potvrzují testy shodu obou rozptylů (Korigovaný F-test) a shodu průměrů (T – test modifikováná šikmost).

Obě metody stanovení arzénu v odpadní vodě vedou ke stejným výsledkům.

Část (C) Párový test

Zadání

V reálných vzorcích podzemních vod stanovovily dvě laboratoře (L1,L2) fotometrickou metodou Cr^{6+}. Na naměřená data aplikujte párový test.

Data:
n = 8

L1	0.56	1.25	0.54	17.6	26.4	1.62	2.05	0.04
L2	0.48	1.40	0.48	18.5	25.8	1.80	1.98	0.05

Řešení:

data									průměr
L1	0.56	1.25	0.54	17.6	26.4	1.62	2.05	0.04	6.2575
L2	0.48	1.40	0.48	18.5	25.8	1.80	1.98	0.05	6.31125
d_i	0.08	-0.15	0.06	-0.9	0.6	-0.18	0.07	-0.01	-0.05375

Program: **ADSTAT 1.25**

Modul: **Jednorozměrná data**

Metoda: **Porovnání dvou výběrů**

Výstup:

(1) Klasické odhady parametrů:

Parametr	L1	L2	celkově
velikost výběru	8	8	16
Průměr	6.2575	6.3112	6.2844
Rozptyl	100.35	99.831	93.419
Šikmost	1.3246	1.2676	1.2962
Špičatost	3.0184	2.8099	2.9147

(2) t – test (párové)

Průměrný rozdíl:	-0.05375
Rozptyl:	0.16321
Počet stupňů volnosti Df1:	7
Tabulkový kvantil t(1-α/2, Df1):	2.3646
t – statistika:	-0.93149

Závěr:
Dvojice se považují za shodné, H0 přijata
Vypočtená hladina významnosti: 0.383

Závěr:

Párový test považuje oba soubory za shodné.

Obě laboratoře poskytují shodné výsledky stanovení Cr^{6+} v podzemních vodách.