UNIVERZITA PARDUBICE
Fakulta chemicko-technologická
Katedra analytické chemie

ANOVA

Semestrální práce

Licenční studium Galileo – Interaktivní statistická analýza dat

Ing. Petra Hlaváčková, Ph.D.
Mendelova univerzita v Brně

Brno 2015

Obsah

1 Úloha 1 Jednofaktorová ANOVA

1.1 Zadání úlohy 1 a vstupní data

V roce 2013 proběhlo na území Školního lesního podniku Masarykův les Křtiny Mendelovy univerzity v Brně dotazníkové šetření zaměřené na uživatele lesních cest a cyklostezek. Tabulka 1.1 uvádí data z dotazníkového šetření. Jedná se o procentní podíl vyplněných dotazníků na čtyřech lokalitách v měsících srpen až říjen. Cílem úlohy je určit jaký vliv mají lokality (faktor A) na počet vyplněných dotazníků v zájmovém území. Použitým programem byl ADSTAT: Analýza rozptylu: ANOVA 1#.

Tab. 1.1 Vstupní data analýzy; n = 12, k = 4

Měsíc/lokalita	Lokalita 1	Lokalita 2	Lokalita 3	Lokalita 4
Srpen	27,94	7,89	42,51	21,66
Září	14,02	9,45	64,94	11,59
Říjen	13,87	7,51	64,74	13,87

1.2 Řešení

Pro řešení byly stanoveny hypotézy:

- **H_0:** Lokality nemají na počet vyplněných dotazníků vliv
- **H_A:** Lokality mají na počet vyplněných dotazníků vliv

Podmínky:

- hladina významnosti α: 0,05
- transformace: ne
- počet úrovní faktoru A, k: 4
- celkový počet n: 12
- velikost: $n_1 = 3$ $n_2 = 3$ $n_3 = 3$ $n_4 = 3$

1.2.1 Průměry a efekty úrovní

Byl proveden výpočet parametrů sloupcových průměrů, celkového průměru, sloupcových efektů, reziduí a diagonálních prvků H_{ii} projekční matice H. Výsledky uvádí tabulka 1.2.

Tab. 1.2 Vypočtené parametry programem ADSTAT

Úroveň	Průměr	Efekt	H_{ii}
1	18,607	-6,392	0,333
2	8,283	-16,715	0,333
3	57,397	32,398	0,333
4	15,707	-9,292	0,333

Celkový průměr: 24,998

Reziduální rozptyl: 65,124

Detekce vlivných bodů – nebyly nalezeny vybočující a odlehlé body

1.2.2 ANOVA tabulka

Dále byla sestavena tabulka 1.3 a proveden F-test významnosti faktoru A.

Hypotézy:

- H_0: Efekty faktoru A jsou nulové
- H_A: Efekty faktoru A nejsou nulové

Kvantil F (1 – α, k – 1, n – k): 4,066

Tab. 1.3 Tabulka ANOVA

Zdroj rozptylu	Stupně volnosti	Součet čtverců	Průměrný čtverec	Testovací kritérium	Závěr	Spočtená hladina významnosti
Mezi úrovněmi	k – 1 = 3	4368,70	1456,20	22,361	H_0 je zamítnuta	0,000
Rezidua	n – k = 8	520,99	65,12			
Celkový	n – 1 = 11	4889,70	444,52			

Závěr: Fisherovo-Snedecorovo testační kritérium F_e = 22,361. Jelikož je tato hodnota vyšší než kvantil $F_{1-0,05}$ (4 – 1, 12 – 4) = 4,066, tj. $F_e > F_{(1-0,05)}$. Spočtená hladina významnosti α = 0,000 je menší než stanovená α = 0,05. Z těchto důvodů se nulová hypotéza zamítá a faktor A je statisticky významný.

1.2.3 Mnohonásobné porovnání

Dále byly testovány lineární kontrasty pro zadané kombinace úrovní H_0: $\mu_i - \mu_j = 0$. K testování byla použita metoda mnohonásobného porovnání tzv. Sheffeho procedura. Výsledky uvádí tabulka 1.4.

Tab. 1.4 Scheffeho mnohonásobné porovnání

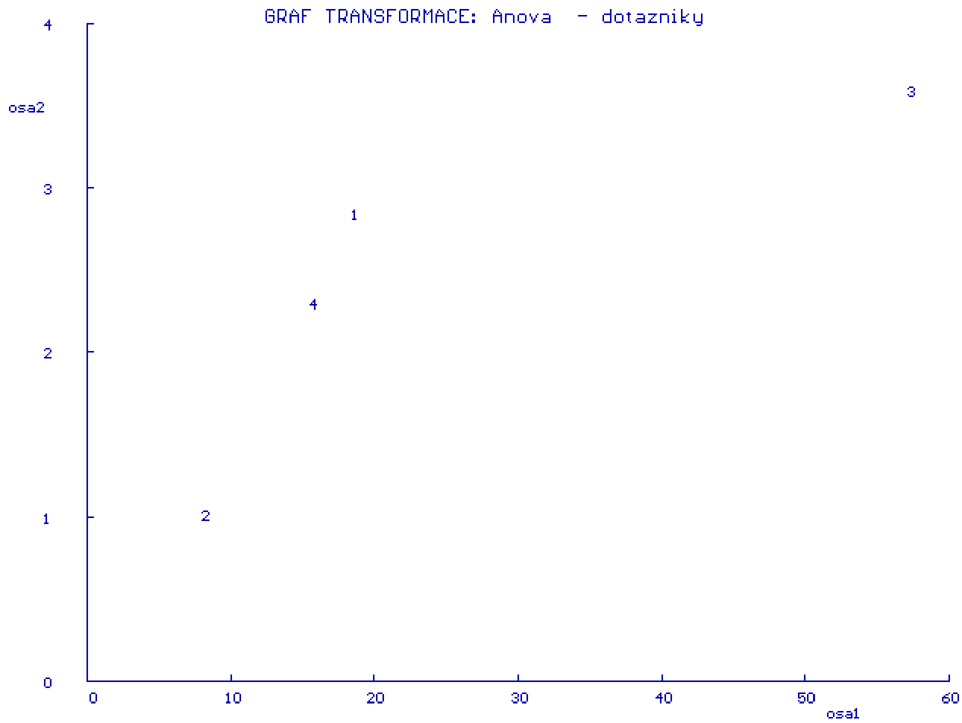
Hypotéza H_0	Průměrný párový rozdíl	Meze konfidenčního intervalu		Závěr
		dolní	horní	
P1 = P2	10,323	-12,69	33,337	akceptována
P1 = P3	-38,790	-61,803	-15,777	zamítnuta
P1 = P4	2,900	-20,113	25,913	akceptována
P2 = P3	-49113,000	-72,127	-26,1	zamítnuta
P2 = P4	-7,423	-30,437	15,59	akceptována
P3 = P4	41,690	18,677	64,703	zamítnuta

Závěr: Ze vzájemného porovnání dvojic faktoru A vyplývá, že nulová hypotéza nebyla přijata pro tři z těchto dvojic, a to P1 = P3, P2 = P3 a P3 = P4. Zde je patrná nerovnost výsledků, velká odlišnost je v případě lokality 3. Zamítnutí hypotézy je zřejmé z korelačních intervalů, kdy u zamítnutých dvojic korelační intervaly neobsahují nulu.

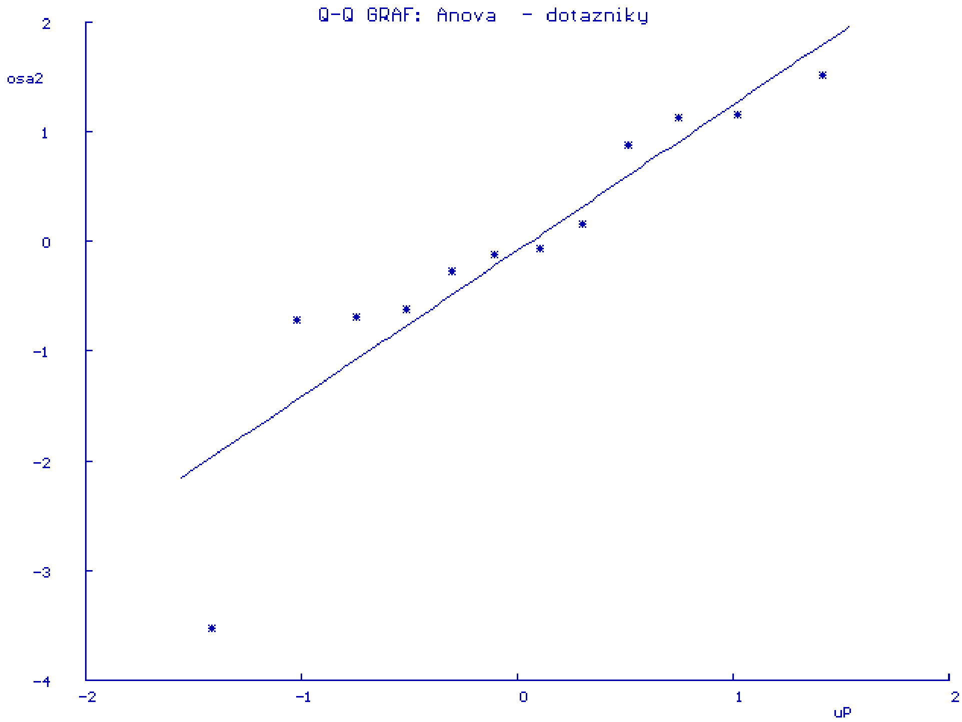
1.2.4 Test transformace a normality

Dále byla provedena zkouška transformace pomocí korelačního koeficientu a grafu transformace (viz obrázek 1.1 a test normality pomocí Q-Q grafu (viz obrázek 1.2), který umožňuje porovnání rozdělení výběru s rozdělením normálním.

Korelační koeficient transformace R dosahuje hodnoty 0,834.



Obr. 1.1 Graf transformace



Obr. 1.2 Q-Q graf

Závěr: Jelikož korelační koeficient dosahuje hodnoty blízké nule a body na obrázku 1.1 nelze proložit přímku, lze konstatovat, že transformace není nutná. Většina bodů dle Q-Q grafu splňuje lineární závislost, lze tedy normalitu přijmout.

1.3 Závěr

Cílem úlohy bylo zjistit, zda mají jednotlivé lokality vliv na počet vyplněných dotazníků. Dle tabulky Anova bylo zjištěno, že nulová hypotéza, tedy, že lokality nemají vliv na počet vyplněných dotazníků, se zamítá. Scheffeho procedura mnohonásobného porovnání ukázala, že statistický rozdíl je především mezi lokalitou 3 a všemi ostatními. Je to dáno tím, že lokalita 3 je lokalitou nejfrekventovanější a je navštěvována jiným okruhem uživatelů stezek. Zkouškou transformace bylo zjištěno, že se transformační koeficient blíží nule a tudíž není nutná transformace dat. Na základě Q-Q grafu lze přijmout předpoklad normálního rozdělení.

2 Úloha 2 Dvoufaktorová ANOVA bez opakování

2.1 Zadání úlohy 2 a vstupní data

V roce 2013 proběhlo na území Školního lesního podniku Masarykův les Křtiny Mendelovy univerzity v Brně dotazníkové šetření zaměřené na uživatele lesních cest a cyklostezek. Šetření probíhalo v měsících červenec až říjen vždy 1 týden v každém měsíci. Tabulka 2.1 uvádí procentní podíl počtu dotazníků v jednotlivých měsících připadající na den v týdnu. Cílem analýzy je zjistit, který z faktorů (faktor A – měsíc, faktor B – den v týdnu) má vliv na počet vyplněných dotazníků. Pro zpracování dat bude využito programu ADSTAT: Analýza rozptylu: ANOVA 2P.

Tab. 2.1 Vstupní data analýzy; n = 28;

Měsíc/den	Po	Út	St	Čt	Pá	So	Ne
Červenec	15,53	15,53	10,75	11,95	12,80	14,85	18,60
Srpen	11,97	14,40	15,82	12,78	10,95	14,20	19,88
Září	10,98	8,54	9,15	13,11	7,62	22,87	27,74
Říjen	12,72	12,72	5,78	7,51	9,25	26,01	26,01

2.2 Řešení

Pro řešení byly stanoveny hypotézy:

- **H_0:** Faktory A (měsíc v roce), B (den v týdnu) a jejich vzájemné interakce nemají vliv na počet vyplněných dotazníků
- **H_A:** Faktory A (měsíc v roce), B (den v týdnu) a jejich vzájemné interakce mají vliv na počet vyplněných dotazníků

Podmínky:

- hladina významnosti α: 0,05
- transformace: ne
- metoda analýzy: nejmenší čtverce

2.2.1 Průměry a efekty úrovní

Byl proveden výpočet parametrů:

- celkový průměr
- reziduální rozptyl
- Tukeyho konstanta
- sloupcové a řádkové průměry
- řádkové (α_i) a sloupcové efekty (β_j)

Celkový průměr: 14,224

Reziduální rozptyl: 17,430

Tukeyho konstanta C: -3,453

Další výsledky jsou uvedeny v tabulce 2.2.

Tab. 2.2 Průměry a efekty srovnání

Faktor A		
úroveň	**průměr**	**efekt**
1	14,287	0,0629
2	14,286	0,0614
3	14,287	0,0629
4	14,037	-0,1871

Faktor B		
úroveň	**průměr**	**efekt**
1	12,8	-1,424
2	12,279	-1,427
3	10,375	-3,849
4	11,337	-2,887
5	9,72	-4,504
6	19,483	5,258
7	23,057	8,833

Z hodnoty směrnice přímky C, která je nenulová, vyplývá, že existuje interakce faktorů a součet čtverců odchylek Tukeyho interakce je k testování.

2.2.2 ANOVA tabulka pro model s Tukeyho interakcí

Byla sestavena tabulka 2.3 a provedeny F-testy významnosti faktorů A, B a AB

Hypotézy:

- H_0: Efekty faktoru A jsou nulové, H_A: Efekty faktoru A nejsou nulové
  Kvantil F(1-α, n-1,mn-m-n): 3,197
- H_0: Efekty faktoru B jsou nulové, H_A: Efekty faktoru B nejsou nulové
  Kvantil F(1-α, n-1,mn-m-n): 2,699
- H_0: Interakce I[1] je nulová, H_A: Interakce I není nulová
  Kvantil F(1-α, n-1,mn-m-n): 4,451

Tab. 2.3 Tabulka ANOVA pro model s Tukeyho interakcí

Zdroj rozptylu	Stupně volnosti	Součet čtverců	Průměrný čtverec	Testovací kritérium	Závěr	Spočtená hladina významnosti
Mezi úrovněmi A	n - 1 = 3	0,327	0,109	0,006	H_0 je akceptována	0,999
Mezi úrovněmi B	n - 1 = 6	612,710	102,120	5,859	H_0 je zamítnuta	0,002
Interakce	1	85,321	85,321	4,895	H_0 je zamítnuta	0,041
Rezidua	mn - m - n = 17	296,300	17,430			
Celkový	mn - 1 = 27	909,340	33,679			

Závěr:

Faktor A – jelikož je Fisher-Snedecorovo testovací kritérium F_e = 0,006, což je nižší hodnota než kvantil $F_{1-0,05}$(4-1, 17) = 3,197 a spočtená hladina významnosti α = 0,999 je vyšší než stanovená α = 0,05, lze konstatovat, že nulová hypotéza H_0: Efekty faktoru A jsou nulové, je přijata a faktor A je statisticky nevýznamný. Počet vyplněných dotazníků není ovlivňován měsícem vyplnění.

Faktor B – hodnota Fisher-Snedecorova testovací kritéria F_e = 5,859 je vyšší než kvantil $F_{1-0,05}$(7-1, 17) = 2,699 a spočtená hladina významnosti α = 0,002 je nižší než stanovená α = 0,05, proto je nulová

[1] I je efekt Tukeyho interakce

hypotéza H_0: Efekty faktoru B jsou nulové, zamítnuta. Z toho vyplývá, že faktor B je statisticky významný. Na počet vyplněných dotazníků má vyliv den v týdnu.

Interakce AB – v případě interakce faktorů A a B nabývá Fisher-Snedecorovo testovací kritérium hodnoty F_e = 4,895, která je vyšší než kvantil $F_{1-0,05}(1, 17)$ = 4,451. Spočtená hladina významnosti α = 0,041 je nižší než stanovená α = 0,05. Z uvedeného vyplývá, že nulová hypotéza H_0: Interakce I je nulová se zamítá a interakce faktorů A a B je statisticky významná. To znamená, že kombinace dne v týdnu a sledovaných měsíců má vliv na počet vyplněných dotazníků.

2.2.3 Transformace

Odhad mocninné transformace: 50,126

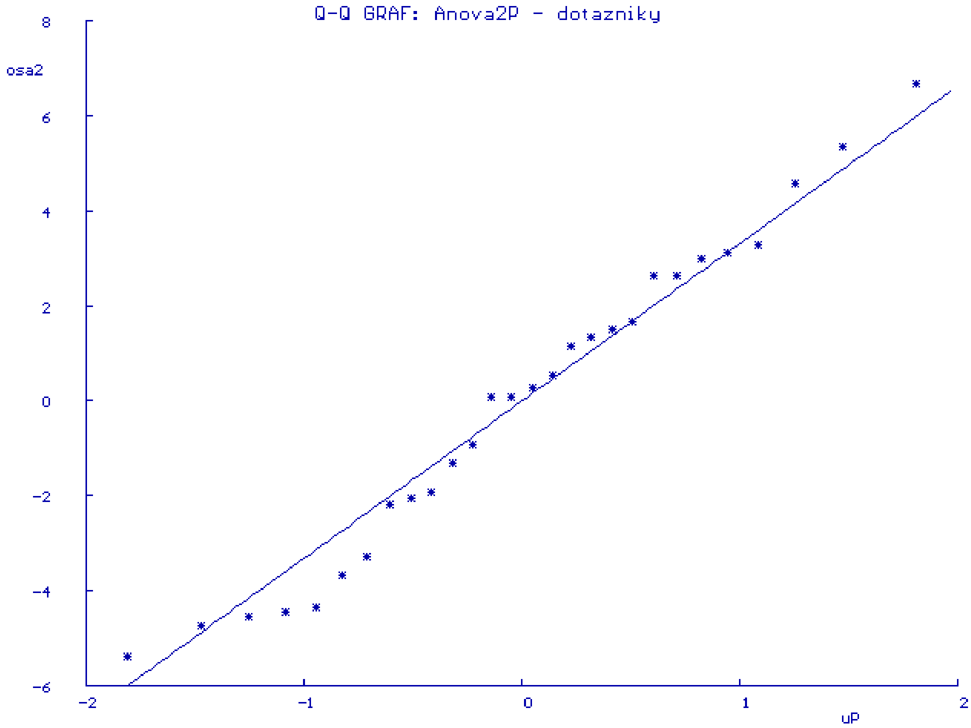
Rozptyl odhadu transformace: 465,620

Akceptovatelný interval: (28,548, 71,704)

Závěr: Z dat je zřejmé, že odhad mocninné transformace leží v akceptovatelném intervalu, z tohoto důvodu není třeba transformovat data mocninnou nebo logaritmickou transformací.

2.2.4 Test normality

Test normality byl proveden pomocí Q-Q grafu (viz obrázek 2.1).



Obr. 2.1 Q-Q graf

Závěr: Body leží přibližně na přímce, tedy splňují lineární závislost, předpoklad normality je přijat.

2.2.5 Závěr

Cílem úlohy bylo zjistit, zda měsíc dotazníkového šetření a den v týdnu ovlivní počet vyplněných dotazníků. Pomocí dvoufaktorové analýzy rozptylu bez opakování bylo zjištěno, že významný vliv má pouze faktor B a interakce faktorů A a B má vliv méně výrazný. Znamená to, že den v týdnu a kombinace měsíce šetření a dne v týdnu ovlivňují počet vyplněných dotazníků.

3 Dvoufaktorová ANOVA s opakováním (vyvážená)

3.1 Zadání úlohy 3 a vstupní data

Průzkum návštěvnosti území Školního lesního podniku Masarykův les Křtiny v roce 2013 se mimo jiné také zaměřil na zjišťování počtu osob užívající lesní cesty a cyklostezky. Měření probíhalo pomocí sčítačů firmy Eco-counter, typ Pyro Box Compact subdodávkou od specializované firmy. Návštěvníci byly zaznamenáváni ve dvou směrech (IN a OUT). Vyhodnocování dat probíhalo v hodinových časových intervalech. V tabulce 3.1 jsou uvedeny hodnoty pro jeden týden za dobu 8 hodin. Cílem analýzy rozptylu s opakováním je zjistit, zda má hodina (faktor A) a den (faktor B) vliv na počet návštěvníků území. Použitým programem byl ADSTAT: Analýza rozptylu 2B.

Tab. 3.1 Vstupní data pro analýzu

Čas	Směr	Pondělí	Úterý	Středa	Čtvrtek	Pátek	Sobota	Neděle
9-10h	IN	11	19	24	62	25	36	28
	OUT	12	21	49	31	47	81	58
10-11h	IN	23	39	37	17	35	63	72
	OUT	27	46	49	58	33	134	165
11-12h	IN	23	37	44	40	35	71	97
	OUT	51	70	80	42	82	129	138
12-13h	IN	22	35	49	27	30	76	104
	OUT	18	66	31	46	45	95	86
13-14h	IN	28	56	36	37	42	72	110
	OUT	29	35	37	34	47	79	92
14-15h	IN	46	35	44	57	34	95	84
	OUT	38	43	53	52	37	93	78
15-16h	IN	37	41	35	26	37	90	139
	OUT	55	74	49	84	70	88	131
16-17h	IN	41	57	56	38	54	120	139
	OUT	81	94	93	74	61	76	134

3.1.1 Řešení

Pro řešení byly stanoveny hypotézy:

- **H_0:** Faktory A (hodina), B (den v týdnu) a jejich vzájemné interakce nemají vliv na počet vyplněných dotazníků
- **H_A:** Faktory A (hodina), B (den v týdnu) a jejich vzájemné interakce mají vliv na počet vyplněných dotazníků

Podmínky:

- hladina významnosti α: 0,05
- transformace: ne
- počet úrovní parametru A, n: 8
- počet úrovní parametru B, m: 7
- počet opakování v jedné buňce, o: 2

3.1.2 Průměry a efekty srovnání

Byl proveden výpočet parametrů:

- celkový průměr
- reziduální rozptyl
- sloupcové a řádkové průměry
- řádkové (α_i) a sloupcové efekty (β_j)

Celkový průměr: 58,643

Reziduální rozptyl: 432,950

Další výsledky jsou uvedeny v tabulce 3.2. Pro každou úroveň faktorů A a B bylo měření dvakrát opakováno.

Tab. 3.2 Průměry a efekty srovnání

Faktor A		
úroveň	**průměr**	**efekt**
1	36,000	-22,643
2	57,000	-1,643
3	67,071	8,429
4	52,143	-6,500
5	52,429	-6,000
6	56,357	-2,286
7	68,286	9,643
8	79,857	21,214

Faktor B		
úroveň	**průměr**	**efekt**
1	33,875	-24,768
2	48,000	-10,643
3	47,875	-10,768
4	45,313	-13,330
5	44,625	-14,018
6	87,375	28,732
7	103,440	44,795

3.1.3 ANOVA tabulka pro model s interakcemi faktorů A a B

Byla sestavena tabulka ANOVA (viz tabulka 3.4) a provedeny F-testy významnosti faktorů A, B a AB.

Hypotézy:

- H_0: Efekty faktoru A jsou nulové, H_A: Efekty faktoru A nejsou nulové
  Kvantil F(1-α, n-1,mn (o-1)): 2,178
- H_0: Efekty faktoru B jsou nulové, H_A: Efekty faktoru B nejsou nulové
  Kvantil F(1-α, m-1, mn (o-1)): 2,266
- H_0: Interakce I[2] je nulová, H_A: Interakce I není nulová
  Kvantil F(1-α, (n-1)(n-1), mn (o-1)): 1,600

Tab. 3.3 Tabulka ANOVA pro model s interakcemi faktorů A a B

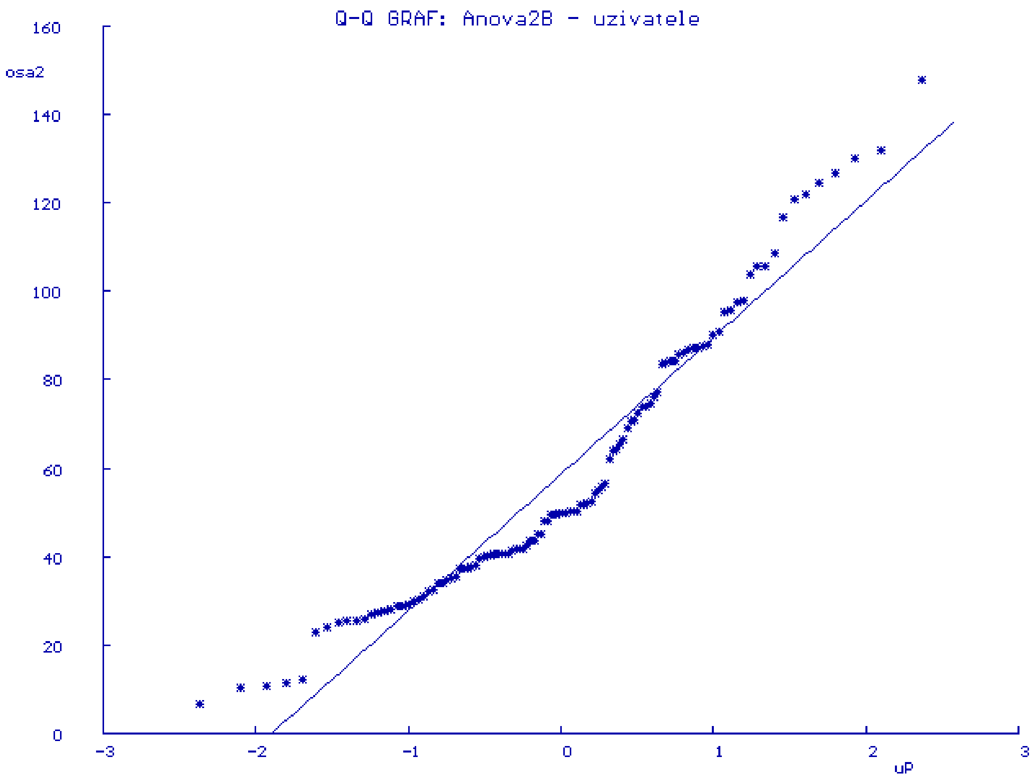
[2] I znamená efekty interakcí A a B dohromady

Zdroj rozptylu	Stupně volnosti	Součet čtverců	Průměrný čtverec	Testovací kritérium	Závěr	Spočtená hladina významnosti
Mezi úrovněmi A	n - 1 = 7	17 018	2 431	5,615	H0 je zamítnuta	0,000
Mezi úrovněmi B	m - 1 = 6	64 783	10 797	24,939	H_0 je zamítnuta	0,000
Interakce	(n-1)(n-1) = 42	11 594	276	0,638	H_0 je akceptována	0,935
Rezidua	mn(o-1) = 49	24 245	433			
Celkový	mno - 1 = 111	117 640	1 060			

Závěr: Analýza dat metodou ANOVA zamítla nulovou hypotézu pro oba faktory A a B. Pro interakci faktorů se nulová hypotéza nezamítá. V případě faktoru A je Fisher-Snedecorovo testovací kritérium F_e = 5,615, což je vyšší hodnota než kvantil $F_{1-0,05}$ = 2,178 a spočtená hladina významnosti α = 0,000 je nižší než stanovená α = 0,05. Nulová hypotéza H_0: Efekty faktoru A jsou nulové, se zamítá a faktor A je statisticky významný. Různé denní hodiny mají vliv na počet návštěvníků území. V případě faktoru B je hodnota testovacího kritéria F_e = 24,939 vyšší než kvantil $F_{1-0,05}$ = 2,266 a spočtená hladina významnosti α = 0,000 je nižší než stanovená α = 0,05, proto je nulová hypotéza H_0: Efekty faktoru B jsou nulové, zamítnuta. Faktor B je statisticky významný, což znamená, že den v týdnu má vliv na návštěvnost zájmového území. Vzájemná interakce mezi denní hodinou a dnem v týdnu však nebyla prokázána, jelikož v tomto případě je testovací kritérium F_e = 1,600 nižší než kvantil $F_{1-0,05}$ = 0,638 a spočtená hladina významnosti α = 0,935 je vyšší než stanovená α = 0,05. Z uvedeného vyplývá, že nulová hypotéza H_0: Interakce I je nulová, se nezamítá a interakce faktorů A a B je statisticky nevýznamná.

3.1.4 Test transformace a normality

Zkouška transformace dle korelačního koeficientu R, který vykázal hodnotu (0,379), která se blíží nule, prokázala, že transformace není nutná. Test normality byl proveden Q-Q grafem (viz obrázek



Obr. 3.1 Q-Q graf

Dle Q-Q grafu je možné konstatovat, že rozdělení je normální.

3.1.5 Závěr

Analýzou rozptylu s opakováním bylo zjištěno, že denní hodina i den v týdnu má vliv na návštěvnost území. Nebyla však prokázána interakce denní hodiny a dne v týdnu a jejich vliv na počet návštěvníků území.

Seznam literatury

Meloun, M., Militký, J. 2012. Kompendium statistického zpracování dat. 3. vyd. Praha: Karolinum Praha. 985 s. ISBN 978-80-246-2196-8.

Meloun, M., Militký, J. 2012. Interaktivní statistická analýza dat. 4. vyd. Praha: Karolinum Praha. 955 s. ISBN 978-80-246-2173-9.