10. ANOVA – Analýza rozptylu

Čas ke studiu kapitoly: 60 minut

Cíl Po prostudování tohoto odstavce budete umět

- porozumět konstrukci *F-poměru*
- rozhodovat se pomocí testu zvaného analýza rozptylu
- zkonstruovat tabulku ANOVA
- provést post hoc analýzu

VÝKLAD

10.1. Úvod

V předcházejících kapitolách jsme se věnovali mimo jiné také jednovýběrovým a dvouvýběrovým testům střední hodnoty. Rozšířením těchto testů je analýza rozptylu neboli ANOVA, která nám umožňuje srovnávat několik středních hodnot nezávislých náhodných výběrů. Na tomto místě je pak třeba zmínit požadavky parametrického testu, který budeme dále užívat (*tabulka ANOVA*). Analýza rozptylu ve své parametrické podobě předpokládá normalitu rozdělní a tzv. homoskedasticitu (identické rozptyly). Pokud tyto podmínky nejsou splněny, je třeba použít neparametrický *Kruskal-Wallisův test*, který je obdobou jednofaktorového třídění v analýze rozptylu (v závěru této lekce bude uveden jen v náznaku). Na rozdíl od parametrického testu však nepředpokládá normalitu rozdělení, jeho nevýhodou je pak menší citlivost.
Analýza rozptylu tak představuje rozšíření možností procedury zvané testování hypotéz.

10.2. Konstrukce F-statistiky

Nechť máme k-náhodných výběrů (tj. výběry z k populací), které jsou na sobě nezávislé.
Nechť tyto náhodné výběry pochází z normálních rozdělení se stejným rozptylem:

$(X_{11}, X_{12}, ..., X_{1n_1}) \rightarrow N(\mu_1, \sigma^2)$

$(X_{21}, X_{22}, ..., X_{2n_2}) \rightarrow N(\mu_2, \sigma^2)$

...

$(X_{k1}, X_{k2}, ..., X_{kn_k}) \rightarrow N(\mu_k, \sigma^2)$, nechť n_i = počet pozorování v i-tém náhodném výběru

---

$$\sum_{i=1}^{k} n_i = N$$

***Formulace problému*:**

Je třeba testovat hypotézu H_0: $\mu_1 = \mu_2 = ... = \mu_k = \mu$
vůči alternativě : H_A: neplatí H_0

Chceme rozhodnout o H_0 na základě jednoho testu. Proto se pokusíme nalézt takovou testovou statistiku, která nejen umožní implementaci H_0 , ale je i citlivá na platnost H_0.

Definujme **totální součet čtverců** (nebo **totální variabilitu**) jako

$SS_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2$, kde $\overline{X}$ je výběrový průměr ze všech pozorovaných hodnot.

Tento totální součet čtverců můžeme snadno rozložit na 2 složky:

$$SS_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2 \Rightarrow SS_{TOTAL} = SS_W + SS_B \ ,$$

kde

SS_W ... vnitřní variabilita $\quad SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2$

přičemž S_i je výběrová směrodatná odchylka *i*-tého náhodného výběru: $S_i = \sqrt{\frac{\sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2}{n_i - 1}}$

a dále $\overline{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$; je výběrový průměr v *i*-tém náhodném výběru.

SS_B ... mezitřídní variabilita $\quad SS_B = \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2$

Zavedeme následující výběrové rozptyly:

vnitřní výběrový rozptyl $\quad S_W^2 = \frac{SS_W}{N - k}$

mezitřídní výběrový rozptyl $\quad S_B^2 = \frac{SS_B}{k - 1}$

Vlastnosti těchto výběrových rozptylů:

1. $ES_W^2 = \frac{1}{N-k} E(\sum_{i=1}^{k} (n_i - 1) S_i^2) = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) E(S_i^2) = \sigma^2$

   neboť $E(S_i^2) = \sigma^2$.
   Tedy vnitřní výběrový rozptyl je nestranným odhadem rozptylu, nezávisle na H_0.

2. Podobně bychom mohli dokázat, že $ES_B^2 = \sigma^2 + \frac{1}{k-1} \sum_{i=1}^{k} n_i (E\overline{X} - E\overline{X}_i)^2$, z čehož bezprostředně vyplývá následující ekvivalence:
   $ES_B^2 = \sigma^2 \Leftrightarrow$ když platí H_0

Položíme $F = \frac{S_B^2}{S_W^2}$

Definice:
Tuto statistiku F nazveme ***F-poměr*** .

Proč je výhodné použít *F-poměr* jako testovou statistiku?

Z výše uvedeného je zřejmé, že pokud platí H_0, *F-poměr* je nějaké náhodné číslo blízké jedničce ... $F \approx 1$. Dále, pokud neplatí H_0, je toto číslo výrazně větší než 1, jak ukazuje vlastnost 2 – výpočet střední hodnoty mezitřídního výběrového rozptylu. Statistika *F-poměr* je tedy citlivá na platnost hypotézy H_0. Abychom ji mohli v dalším průběhu testu použít jako testovou statistiku (a tím i nulové rozdělení), musíme determinovat její statistické chování, tedy určit její rozdělení pravděpodobnosti.

Víme, že $\frac{S_w^2}{\sigma^2} \cdot (N-k) = \sum_{i=1}^{k} \frac{(n_i - 1) S_i^2}{\sigma^2} \rightarrow \chi^2(N-k)$,

protože $\frac{(n_i - 1) S_i^2}{\sigma^2} \rightarrow \chi^2(n_i - 1)$, a dále je známo, že součet náhodných veličin $\chi^2(n_i - 1)$ je opět náhodnou veličinou stejného typu, s počtem stupňů volnosti daným součtem stupňů volnosti sčítancových veličin.
Podobnou úvahou lze prokázat, že pokud platí H_0, potom:

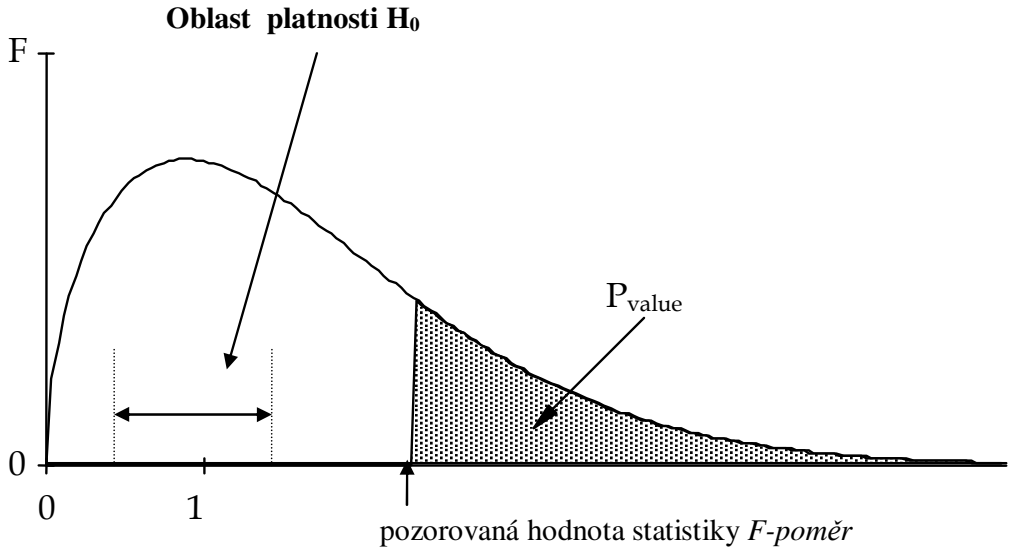
$$\frac{S_B^2}{\sigma^2} \cdot (k-1) = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2 = \sum_{i=1}^{k} \left(\frac{\overline{X}_i - \overline{X}}{\frac{\sigma}{\sqrt{n_i}}}\right)^2 \rightarrow \chi^2(k-1)$$

Pokud tedy platí H_0, potom víme (ze znalostí o Fisherově-Snedecorově rozdělení), že následující podíl:

$$\frac{\frac{S_B^2}{\sigma^2} \cdot (k-1)}{k-1} \Big/ \frac{\frac{S_w^2}{\sigma^2} \cdot (N-k)}{N-k} = \frac{S_B^2}{S_W^2} = F_{k-1, N-k}$$

musí mít nutně *F rozdělení* o *(k-1)* a *(N-k)* stupních volnosti.

Pokud známe statistické chování F-poměru, lze to využít pro účely posouzení a rozhodnutí výše uvedeného problému v podobě H_0. Následující obrázek ilustruje použití *F-poměru* pro účely rozhodování o platnosti hypotézy H_0.



10.3. Tabulka ANOVA

Jednotlivé mezivýsledky, prováděné v průběhu analýzy rozptylu, jsou průběžně a systematicky zaznamenávány v tabulce ANOVA:

Zdroj proměnlivosti	Variabilita	Stupně volnosti	Odpovídající druh rozptylu	Testová stat. *F-poměr*	P-value
totální	$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2$	$N-1$			
mezitřídní	$SS_B = \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2$	$k-1$	$S_B^2 = \frac{SS_B}{k-1}$	$F = \frac{S_B^2}{S_W^2}$	viz. definice
vnitřní	$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$	$N-k$	$S_W^2 = \frac{SS_W}{N-k}$		

Tabulka analýzy rozptylu - ANOVA

Velké hodnoty *F-poměru* budou mít za následek malé hodnoty p_{value}, což znamená zamítnutí H_0. *F-poměru* bude velký, pokud vnitřní variabilita tvoří zanedbatelnou část totální variability a ekvivalentně, pokud mezitřídní variabilita tvoří významnou část totální variability.

10.4. Řešené příklady

Pro ilustraci statistického chování *F-poměru* uvažujme tři datové soubory. Ve všech jsou stejné výběrové průměry v rámci *i*-té populace, avšak rozptyly se liší. Pokud vnitřní výběrový rozptyl je malý, *F-poměr* je velký, pokud je naopak velký, *F-poměr* je malý. Soubory ilustrují tři případy: malý vnitřní výběrový rozptyl, normální a velký.

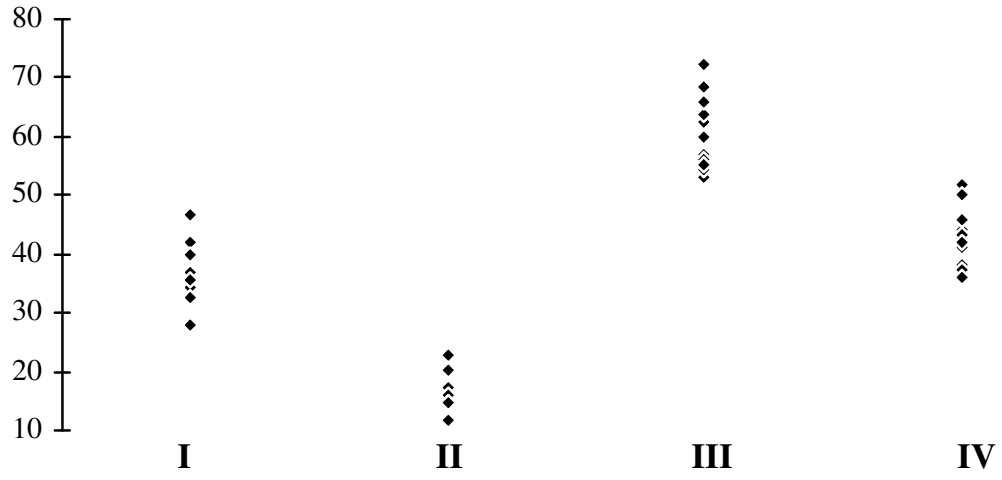
Příklad 1: **Malý vnitřní výběrový rozptyl**

Populace	I	II	III	IV
Data	42 34.5 32.5 40 46.5 28 37 35.5	17.5 12 16 15 20.5 23 15	68.5 72 53 64 57 56 54.5 62.5 63.5 60 66 55	38 44 52 50 43.5 41 42 46 37.5 36
Rozsah výběru	8	7	12	10
Výběrové průměry	37	17	61	43
Výběrové směrodatné odchylky	5.78	3.71	6.06	5.27

Tabulka ANOVA

	Počet stupňů volnosti	Variabilita	Odpovídající výb. rozptyl	*F-poměr*
totální	36	9872.7027		
mezitřídní	3	8902.7027	2967.57	100.96
vnitřní	33	970	29.39	

P-value = 0.0000



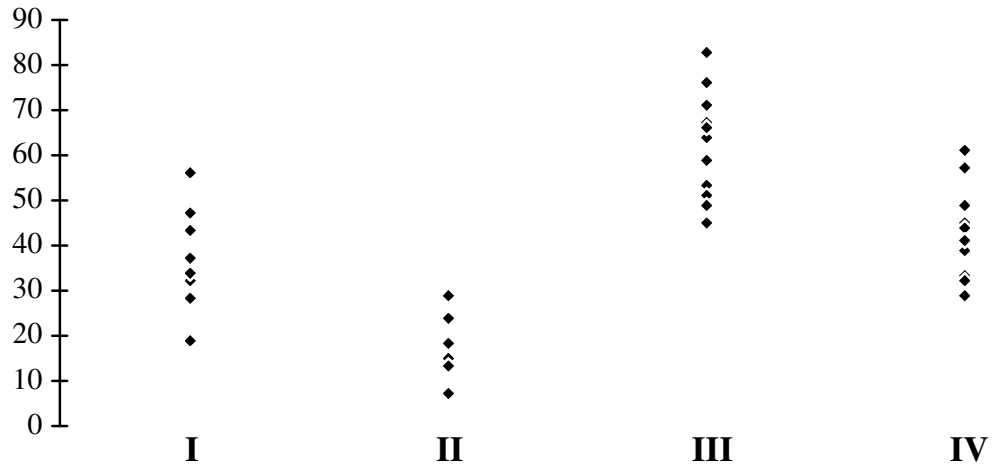
Příklad 2: **Normální vnitřní výběrový rozptyl**

Populace	I	II	III	IV
Data	47 32 28 43 56 19 37 34	18 7 15 13 24 29 13	76 83 45 67 53 51 48 64 66 59 71 49	33 45 61 57 44 39 41 49 32 29
Rozsah výběru	8	7	12	10
Výběrové průměry	37	17	61	43
Výběrové směrodatné odchylky	11.56	7.42	12.12	10.53

Tabulka ANOVA

	Počet stupňů volnosti	Variabilita	Odpovídající výb. rozptyl	*F-poměr*
totální	36	12782.7027		
mezitřídní	3	8902.7027	2967.57	25.24
vnitřní	33	3880	117.58	

P-value = 0.0000



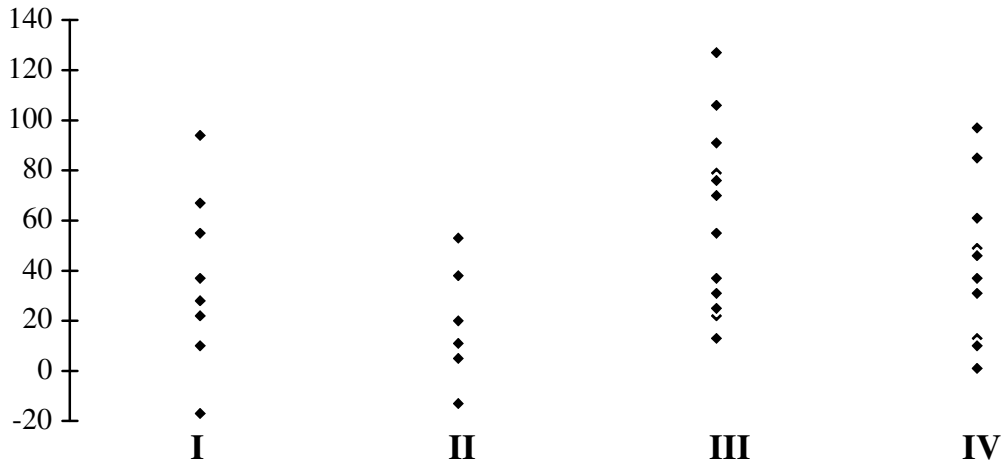
Příklad 3: **Velký vnitřní výběrový rozptyl**

Populace	I	II	III	IV
Data	67 22 10 55 94 -17 37 28	20 -13 11 5 38 53 5	106 127 13 79 37 31 22 70 76 55 91 25	13 49 97 85 46 31 37 61 10 1
Rozsah výběru	8	7	12	10
Výběrové průměry	37	17	61	43
Výběrové směrodatné odchylky	34.69	22.25	36.36	31.59

Tabulka ANOVA

	Počet stupňů volnosti	Variabilita	Odpovídající výb. rozptyl	*F-poměr*
totální	36	43822.7027		
mezitřídní	3	8902.7027	2967.57	2.804
vnitřní	33	34920	1058.18	

P-value = 0.0549



10.5. Post Hoc analýza

Předchozí analýza poukázala na to, že velký *F-poměr* indikuje existenci významných změn mezi populačními výběrovými průměry. Naše analýza by ale byla nekompletní, pokud bychom neidentifikovali, které z populací signalizují významnou odchylku výběrového průměru. Tento další proces se nazývá **post hoc** analýza a spočívá v porovnávání výběrových průměrů všech dvojic populací.
Pro tato vícenásobná porovnávání existuje několik metod. V rámci tohoto výkladu se omezíme jen na tu nejjednodušší z nich, tzv. **LSD-metodu** (znamená zkratku výrazu Lest Significant Difference). Tato metoda spočívá v aplikaci dvouvýběrového t-testu pro každý pár výběrových průměrů. Místo standardního dvouvýběrového Studentova t-testu však použijeme poněkud upravený t-test, založený na LSD statistice:
Pro *i*-tý a *j*-tý výběr definujeme následující testovou statistiku $(LSD)_{i,j}$:

$$(LSD)_{i,j} = \frac{\overline{X}_i - \overline{X}_j}{S_W \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \rightarrow t_{N-k}$$

kde $S_W = \sqrt{S_W^2} = \sqrt{\frac{SS_W}{N-k}}$.
Snadno lze zdůvodnit, že tato statistika má Studentovo rozdělení s *N-k* stupni volnosti.

LSD metoda je ilustrována pro tři předchozí příklady:

Příklad 1: Malý vnitřní výběrový rozptyl

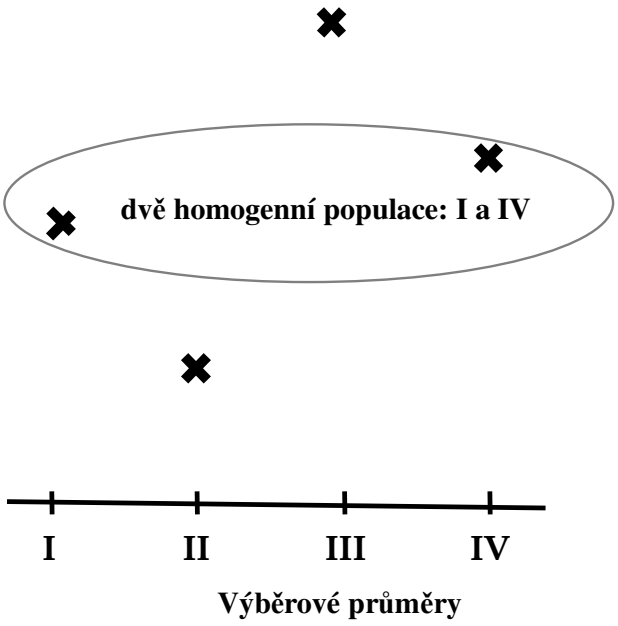
Provedeme výpočet statistiky $(LSD)_{i,j}$ pro všechny uvažované dvojice daných čtyř populací a hodnoty zaznamenáme do následující tabulky:

Rozsahy		8	7	12	10
výběru		I	II	III	IV
8	I	0	-7.128	9.698	2.333
7	II	7.128	0	17.064	9.731
12	III	-9.698	-17.064	0	-7.754
10	IV	-2.333	-9.731	7.7541	0

V tomto případě existuje velmi silná empirická výpověď o rozdílech mezi všemi populacemi, pouze při porovnání populací I a IV výpověď není tak silná.

Příklad 2: Normální vnitřní výběrový rozptyl

Rozsahy		8	7	12	10
výběrů		I	II	III	IV
8	I	0	-3.564	4.849	1.167
7	II	3.564	0	8.532	4.8656
12	III	-4.849	-8.532	0	-3.877
10	IV	-1.167	-4.866	3.877	0



V tomto případě, ačkoliv výběrové průměry jsou stejné, neexistuje empirická výpověď o rozdílu mezi výběrovými průměry populací I a IV. Takže můžeme v podstatě existující 4 populace rozdělit na 3 skupiny: první sdružuje populace I a IV, druhou tvoří populace II a třetí populace III.

Příklad 3: Velký vnitřní výběrový rozptyl

Jelikož *F-poměr* je v tomto příkladě velmi malý, za normálních okolností bychom tento příklad uzavřeli tím, že nezamítáme nulovou hypotézu o rovnosti středních hodnot populací, čímž by analýza skončila, neboť všechny populace jsou homogenní, co do rovnosti středních hodnot. Pokud přesto provedeme výpočet hodnot tabulky $(LSD)_{i,j}$, dostaneme:

Rozsahy		8	7	12	10
výběrů		I	II	III	IV
8	I	0	-1.188	1.616	0.389
7	II	1.188	0	2.844	1.622
12	III	-1.616	-2.844	0	-1.292
10	IV	-0.389	-1.622	1.292	0

V tomto hypotetickém případě vidíme významný rozdíl, který signalizuje malé **P-value** a tedy zamítnutí testu o rovnosti výběrových průměrů, mezi populacemi II a III. Jelikož však celkový *F-poměr* byl příliš malý, tento rozdíl by byl za normálních okolností přehlédnut a my bychom uzavřeli test tím, že neexistují žádné významné rozdíly mezi danými čtyřmi populacemi. Za těchto okolností můžeme tento rozdíl považovat za falešně významný.

Poznámka:
Existují i jiné testy, nežli LSD metoda, které umožňují podobná vícenásobná porovnávání, čili post hoc analýzu. Byly vyvinuty i flexibilnější metody, které jsou dostupné prostřednictvím vyspělého softwaru. Patří sem například Duncanův test, Tukeyův test pro významné rozdíly, Scheffé test a Bonferoni test. Detaily k nim zde nebudou probírány, ale všechny jsou založeny na podobné rozhodovací strategii, založené na stanovení kritického rozdílu požadovaného pro určení toho, zda dva výběrové průměry z několika populací se liší. V mnoha případech jsou tyto testy mnohem efektivnější, než LSD metoda, pro účely nalezení podskupin původních populací, které jsou homogenní co do rovnosti výběrových průměrů.

10.6. Kruskal-Wallisův test

Předchozí postup ANOVA, využívající pro rozhodování popsaný *F-poměr* je velmi citlivý na předpoklad o normalitě rozdělení původních náhodných výběrů. Pro případy, kdy tomuto předpokladu nelze úplně vyhovět, existuje Kruskal Wallisův pořadový test. Neuvádím zde detaily tohoto testu, jen základní myšlenkový postup. Tento test je založen na pořadí původních datových hodnot a provádí analýzu rozptylu takto uspořádaných hodnot. Pro výše uvedený příklad 3 přináší následující tabulka pořadí všech zaznamenaných hodnot:

Populace	I	II	III	IV
	28	11	36	9.5
Pořadí	12.5	2	37	23
původních	6.5	8	9.5	35
hodnot	25.5	4.5	31	32
	34	21	19	22
	1	24	16.5	16.5
	19	4.5	12.5	19
	15		29	27
			30	6.5
			25.5	3
			33	
			14	
Rozsah výběru	8	7	12	10
Průměrné pořadí	17.6875	10.7143	24.4167	19.35
Směrodatná odch.	11.1674	8.5919	9.6668	10.6538

Testová statistika je modifikací dříve uvedeného *F-poměru* pro takto uspořádané hodnoty. Pozorovaná hodnota této, tzv. K-W testové statistiky a příslušná hodnota p-value jsou v daném případě následující:

K-W testová statistika = 7.24325 p-value = 0.0645

P-value pro tuto K-W testovou statistiku je o něco větší, než dává *F-poměr*, ale závěry jsou v obou případech stejné. Nulová hypotéza není zamítnuta.

Shrnutí pojmů

Rozšířením dvouvýběrových testů pro střední hodnoty je **analýza rozptylu** neboli **ANOVA,** která umožňuje srovnávat několik středních hodnot nezávislých náhodných výběrů.
Testovou statistikou je při analýze rozptylu ***F-poměr*** , který byl odvozen na základě analýzy variability vstupních datových souborů. Statistika *F-poměr* je citlivá na platnost hypotézy H_0, která je formulována jako rovnost středních hodnot zkoumaných náhodných výběrů.
Jednotlivé mezivýsledky, prováděné v průběhu analýzy rozptylu, jsou průběžně a systematicky zaznamenávány v **tabulce ANOVA.** Druhým krokem při analýze rozptylu je **post hoc** analýza, která spočívá v porovnávání výběrových průměrů všech dvojic populací s cílem vybrat homogenní (srovnatelné) populace. Kritériem pro zařazení do homogenních skupin může být například **LSD-statistika.**
Popsaný postup ANOVA, využívající pro rozhodování *F-poměr*, je citlivý na předpoklad o normalitě rozdělení původních náhodných výběrů. Pro případy, kdy tomuto předpokladu nelze úplně vyhovět, existuje **Kruskal Wallisův** pořadový test**.**

Otázky

1. Popište konstrukci a stochastické chování statistiky *F-poměr*
2. Co je to vnitřní a mezitřídní výběrový rozptyl ?
3. Jaký je obvyklý výstup z analýzy rozptylu ?
4. Co je to post hoc analýza a LSD-statistika ?

Úlohy k řešení

Př. 1:

Byl proveden průzkum závislosti příjmu na vzdělání lidí. V tabulce jsou uvedeny příjmy v tisících Kč u náhodně vybraných sedmi mužů na každé úrovni vzdělání. (Z - základní, S - středoškolské, V - vysokoškolské).

	Z	S	V
1	10.9	8.9	11.2
2	9.8	10.3	9.7
3	6.4	7.5	15.8
4	4.3	6.9	8.9
5	7.5	14.1	12.2
6	12.3	9.3	17.5
7	5.1	12.5	10.1

Proveďte jednoduché třídění a rozhodněte, zda vzdělání má vliv na příjem.

{p-value = 0.057}

Př. 2:

Z velkého souboru domácnosti bylo náhodně vybráno 5 jednočlenných domácnosti, 8 dvoučlenných, 10 tříčlenných, 10 čtyřčlenných a 7 pětičlenných domácnosti, dohromady tedy 40 domácnosti a byly sledovány jejich měsíční výdaje za potraviny a nápoje připadající na jednoho člena domácnosti (v Kč). Ověřte pomocí analýzy rozptylu, zda se měsíční výdaje za potraviny (na osobu) liší podle počtu členů domácnosti. {Použijte vhodný programový balík}

	Výdaje na jednoho člena domácnosti (v Kč)				
Počet čl. domácnosti	**1**	**2**	**3**	**4**	**5**
	3.440	2.350	2.529	2.137	2.062
	4.044	3.031	2.325	2.201	2.239
	4.014	2.143	2.731	2.786	2.448
	3.776	2.236	2.313	2.132	2.137
	3.672	2.800	2.303	2.223	2.032
		2.901	2.565	2.433	2.101
		2.656	2.777	2.224	2.121
		2.878	2.899	2.763	
			2.755	2.232	
			3.254	2.661	

Př. 3:

Při rozboru efektivnosti bytové výstavby byly u náhodně vybraných dokončených mimopražských bytů třech typů X,Y a Z zaznamenány náklady na $1m^2$ bytové plochy. Výsledky šetření:

Typ X (Kč)	6 825	7 100	7 555	6 890	7 175	7 300	6 905	
Typ Y (Kč)	6 405	6 570	6 325	6 895	6 905	6 550	6 750	6 965
Typ Z (Kč)	7 050	7 355	6 810	6 910	6 700			

Pokuste se prokázat existenci rozdílů v nákladech mezi jednotlivými typy bytů.