13 JEDNOFAKTOROVÁ ANOVA

Čas ke studiu kapitoly: 120 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:

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. My se budeme zabývat tzv. jednofaktorovou ANOVOU (ANOVOU při jednoduchém třídění).

Na tomto místě je pak třeba zmínit požadavky parametrického testu, který budeme dále užívat. Analýza rozptylu (ANOVA, ANalysis Of VAriance) 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. Na rozdíl od parametrického testu nepředpokládá normalitu rozdělení, jeho nevýhodou je pak menší citlivost.

Analýza rozptylu tedy představuje rozšíření možností procedury zvané testování hypotéz o střední hodnotě (jde o vícevýběrový test střední hodnoty).

Pro ilustraci si uveďme motivační příklad, jenž nás provede touto kapitolou.

Naším úkolem je porovnat úspěšnost absolventů gymnázii, SPŠ a odborných učilišť s maturitou (OU) u přijímací zkoušky z matematiky.

Protože tyto typy škol reprezentují studenti různých škol (není gymnázium jako gymnázium…), s různými studijními výsledky a různým nadáním na matematiku, a také vlivem dalších neodstranitelných znaků, bodové hodnocení zástupců jednotlivých typů škol značně kolísá.

Dosažené výsledky náhodně vybraných patnácti studentů jsou uvedeny v následující tabulce.

Gymnázium	SPŠ	OU
55	54	47
54	50	53
58	51	49
61	51	50
52	49	46

Příklad je specifický v tom, že počet pozorování v jednotlivých výběrech je totožný, což nemusí být splněno. V závěru kapitoly proto provedeme **zobecnění výsledků pro případ s různým počtem pozorování v jednotlivých výběrech (třídách)**.

13.1 Jednofaktorová ANOVA pro stejný počet pozorování v jednotlivých výběrech

13.1.1 Rozptyl mezi třídami

Nejdříve se budeme zabývat otázkou, zda výsledky studentů se opravdu liší podle toho jaký typ střední školy absolvovali. Neboli – jsou průměry jednotlivých výběrů (**tříd**) rozdílné vlivem různých středních hodnot příslušných populací, nebo lze rozdíly mezi průměry přičíst na vrub náhodnému kolísání?

Je třeba testovat hypotézu H_0: $\mu_1 = \mu_2 = \mu_3$,

kde μ_1 je střední bodové hodnocení přijímacích zkoušek z matematiky absolventů gymnázia,
μ_2 je střední bodové hodnocení přijímacích zkoušek z matematiky absolventů SPŠ,
μ_3 je střední bodové hodnocení přijímacích zkoušek z matematiky absolventů OU.

vůči alternativě: H_A: $\overline{H_0}$ (neplatí H_0)

Test této hypotézy vyžaduje v první řadě kvantifikaci rozdílů výběrových průměrů $\overline{X_i}$. Vhodným kvantifikátorem je jejich rozptyl, tzv. **rozptyl mezi třídami** $\left(S_B^2\right)$.

$$S_B^2 = \frac{1}{k-1} \cdot \sum_{i=1}^{k} \left(\overline{X_i} - \overline{X}\right)^2 ,$$

kde k je počet tříd (v našem případě je k=3),
n je počet pozorování v jednotlivých třídách (v našem případě je n=5),
$N = \sum_{i=1}^{k} n$ je celkový počet pozorování
$\overline{X_i}$ je průměr i-tého náhodného výběru (i-té třídy),
$\overline{X}$ je celkový průměr (průměr všech hodnot, což je také průměr průměrů $\overline{X_i}$)

$$\overline{X_i} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n} \qquad \overline{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}}{\sum_{i=1}^{k} n} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}}{N} = \frac{\sum_{i=1}^{k} \overline{X_i}}{k}$$

V souvislosti s ANOVOU se mnohdy setkáváme s pojmem **mezitřídní součet čtverců** (neboli **mezitřídní variabilita**) $\left(SS_B\right)$.

$$SS_B = n \cdot \sum_{i=1}^{k} \left(\overline{X_i} - \overline{X}\right)^2$$

V našem případě:

	Gymnázium	SPŠ	OU	
	55	54	47	
	54	50	53	
	58	51	49	
	61	51	50	
	52	49	46	
$\overline{X_i}$	56	51	49	$\overline{\overline{X}} = 52$
$\overline{X_i} - \overline{\overline{X}}$	4	-1	-3	$\sum_{i=1}^{3}\left(\overline{X_i} - \overline{\overline{X}}\right) = 0$
$\left(\overline{X_i} - \overline{\overline{X}}\right)^2$	16	1	9	$\sum_{i=1}^{3}\left(\overline{X_i} - \overline{\overline{X}}\right)^2 = 26$

$$S_B^2 = \frac{1}{3-1} \cdot \sum_{i=1}^{3}\left(\overline{X_i} - \overline{\overline{X}}\right)^2 = \frac{26}{2} = 13{,}0$$

$$SS_B = 5 \cdot \sum_{i=1}^{k}\left(\overline{X_i} - \overline{\overline{X}}\right)^2 = 130{,}0$$

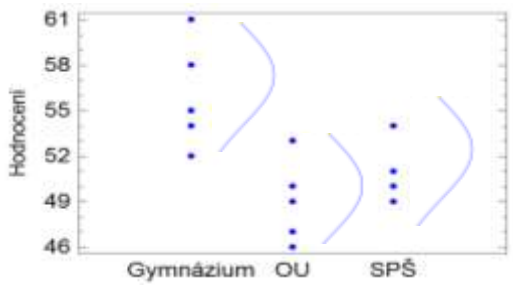
13.1.2 Rozptyl uvnitř tříd

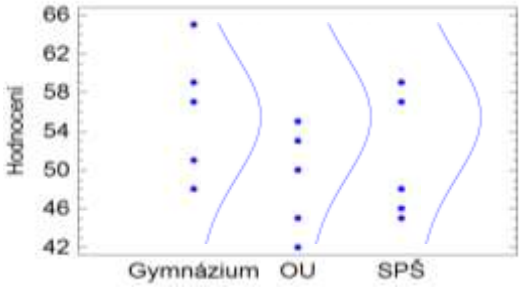
Rozptyl mezi třídami (typy škol) však neposkytuje dostatečnou informaci, neboť nepostihuje kolísání v jednotlivých výběrech.

Pro ujasnění si problému srovnejte údaje ve dvou následujících tabulkách – první z nich uvádí bodové hodnocení náhodně vybraných studentů, druhá taktéž, avšak výsledky ve druhé tabulce vykazují značné kolísání v rámci jednotlivých typů škol. Rozptyly mezi třídami jsou pro oba případy totožné !!!

	Gymnázium	SPŠ	OU	
	55	54	47	
	54	50	53	
	58	51	49	
	61	51	50	
	52	49	46	
$\overline{X_i}$	56	51	49	$\overline{\overline{X}} = 52$
$\overline{X_i} - \overline{\overline{X}}$	4	-1	-3	$\sum_{i=1}^{3}\left(\overline{X_i} - \overline{\overline{X}}\right) = 0$
$\left(\overline{X_i} - \overline{\overline{X}}\right)^2$	16	1	9	$\sum_{i=1}^{3}\left(\overline{X_i} - \overline{\overline{X}}\right)^2 = 26$

	Gymnázium	SPŠ	OU	
	48	57	50	
	57	59	42	
	65	48	53	
	59	46	45	
	51	45	55	
$\overline{X_i}$	56	51	49	$\overline{\overline{X}} = 52$
$\overline{X_i} - \overline{\overline{X}}$	4	-1	-3	$\sum_{i=1}^{3}\left(\overline{X_i} - \overline{\overline{X}}\right) = 0$
$\left(\overline{X_i} - \overline{\overline{X}}\right)^2$	16	1	9	$\sum_{i=1}^{3}\left(\overline{X_i} - \overline{\overline{X}}\right)^2 = 26$





Na výše uvedených obrázcích vidíme, že výsledky studentů uvedené ve druhé tabulce jsou natolik nestálé, že všechny tři výběry lze získat z jedné populace.

Jak můžeme měřit kolísání uvnitř tříd? Je zřejmé, že vhodným měřítkem bude rozptýlenost pozorovaných hodnot v rámci jednotlivých výběrů. Mluvíme o **rozptylu uvnitř tříd** $\left(S_W^2\right)$.

$$S_W^2 = \sum_{i=1}^{k} \frac{\sum_{j=1}^{n}(X_{ij} - \overline{X_i})^2}{n-1} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X_i})^2}{N-k} = \frac{\sum_{i=1}^{k}(n-1)S_i^2}{N-k},$$

kde S_i^2 je výběrový rozptyl *i*-tého náhodného výběru (i-té třídy): $S_i^2 = \frac{\sum_{i=1}^{n}\left(X_{ij} - \overline{X_i}\right)^2}{n-1}$

Obdobně jako u mezitřídního srovnávání, používáme pojem **vnitřní součet čtverců** (neboli **vnitřní variabilita**) $\left(SS_W\right)$.

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X_i})^2$$

Součet mezitřídní a vnitřní variability označujeme jako **celkový součet čtverců** (neboli **celková variabilita**) $\left(SS_{TOTAL}^2\right)$.

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

V našem případě:

$$S_W^2 = \sum_{i=1}^{k} \frac{\sum_{j=1}^{n}(X_{ij} - \overline{X_i})^2}{n-1} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X_i})^2}{N-k} =$$

$$= \frac{(55-56)^2 + (54-56)^2 + \ldots + (54-51)^2 + \ldots + (47-49)^2 + \ldots + (46-49)^2}{15-3} = 7{,}8$$

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X_i})^2 = 94{,}0$$

$$SS_{TOTAL} = SS_W + SS_B = 130{,}0 + 94{,}0 = 224{,}0$$

13.1.3 Testovací kritérium F-poměr (F-ratio)

Položme si nyní zásadní otázku. Je rozptyl mezi třídami $\left(S_B^2\right)$ dostatečně velký vzhledem k rozptylu uvnitř tříd $\left(S_W^2\right)$? Neboli je poměr $\frac{S_B^2}{S_W^2}$ velký?

Běžně se zkoumá nepatrně modifikovaný poměr, který se nazývá **F-poměr**, na počest známého anglického statistika Ronalda Fishera (1890-1962):

$$F - ratio = \frac{n \cdot S_B^2}{S_W^2}$$

n je v čitateli uvedeno proto, aby se hodnota F pohybovala kolem 1.

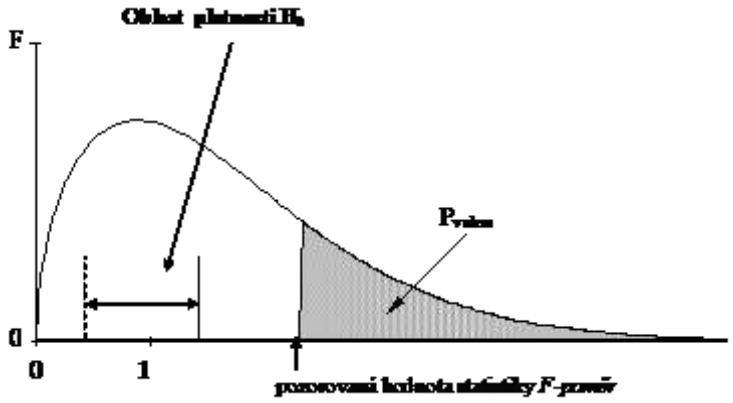
Není-li H_0 pravdivá (střední hodnoty nejsou stejné), pak $n \cdot S_B^2$ bude relativně velké vůči S_W^2 a poměr F bude mnohem větší než 1. Čím větší je F, tím méně je H_0 pravděpodobná.

F-poměr má **Fisher-Snedecorovo rozdělení** s počtem stupňů volnosti

pro čitatele **(k-1)** a pro jmenovatele **(N-k)**

Abychom test mohli dokončit, zbývá nám stanovit si způsob výpočtu p-value pro ANOVU. Z definice p-value (viz. obrázek) je zřejmé, že:

$$p - value = 1 - F(F - ratio)$$



Dokončeme nyní řešení našeho příkladu. Určili jsme, že:

$$S_B^2 = 13{,}0\,, \qquad S_W^2 = 7{,}8\,, \qquad \text{n=5}$$

pak

$$x_{OBS} = F = \frac{n \cdot S_B^2}{S_W^2} = \frac{5 \cdot 13{,}0}{7{,}8} = 8{,}3$$

F-poměr má Fisher-Snedecorovo rozdělení s 2 (=3-1) stupni volnosti pro čitatele a 12 (=3.(5-1)) stupni volnosti pro jmenovatele. V tabulkách pro Fisher-Snedecorovo rozdělení (Tabulka 4) najdeme, že:

$$0{,}990 < \mathrm{F}(8{,}3) < 0{,}999 \qquad \Rightarrow \qquad 0{,}001 < 1 - \mathrm{F}(8{,}3) < 0{,}010$$

a proto

$$0{,}001 < p - value < 0{,}010$$

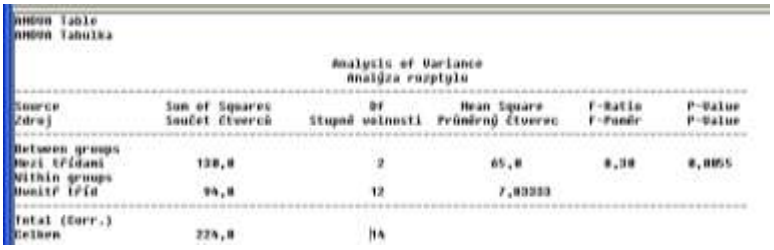
Na 5% -ní hladině významnosti tedy můžeme nulovou hypotézu zamítnout a tvrdit, že střední hodnoty bodového hodnocení u přijímacích zkoušek z matematiky nejsou pro absolventy uvedených tří typů SŠ stejné, tzn., že úspěch u přijímacích zkoušek z matematiky závisí na typu absolvované SŠ.

13.1.4 Tabulka ANOVA

Výpočty, které jsme dosud prováděli ručně, lze mnohem snadněji sledovat, uspořádáme-li je do tabulky standardního tvaru. Tyto tabulky nazýváme tabulky ANOVA.

Uvedeme si tabulku ANOVA obecně pro stejné rozsahy výběrů a konkrétní tabulku vygenerovanou pro náš příklad v software Statgraphics (doplnili jsme české popisy).

Zdroj proměnlivosti	Součet čtverců	Stupně volnosti	Průměrný čtverec	Testová stat. F-poměr	P-value
Mezitřídní (faktor)	$SS_B = n \cdot \sum_{i=1}^{k} (\overline{X}_i - \overline{X})^2$	$k-1$	$MS_B = \frac{SS_B}{k-1}$		
Vnitřní (reziduální)	$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2$	$N-k$	$MS_W = \frac{SS_W}{N-k}$	$F-ratio = \frac{MS_B}{MS_W}$	$1-F(F-ratio)$
Celkový	$SS_{TOTAL} = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \overline{X})^2$	$N-1$			


ANOVA Table
ANOVA Tabulka

Analysis of Variance
Analýza rozptylu

Source Zdroj	Sum of Squares Součet čtverců	Df Stupně volnosti	Mean Square Průměrný čtverec	F-Ratio F-Poměr	P-Value P-Value
Between groups Mezi třídami	130,0	2	65,0	8,38	0,0055
Within groups Uvnitř tříd	94,0	12	7,83333		
Total (Corr.) Celkem	224,0	14			

V prvním sloupci uvádíme možné **zdroje proměnlivosti** jednotlivých pozorování. Vzhledem k tomu, že se zabýváme jednofaktorovou ANOVOU, je zde uveden jeden hlavní zdroj (**faktor**, což byl v našem případě typ SŠ) způsobující rozdíly mezi jednotlivými třídami a ostatní zdroje jsou zahrnuty do druhé skupiny označené **reziduální**. Reziduální zdroje nebyly identifikovány a jsou tudíž příčinou náhodného kolísání uvnitř tříd.

Vydělíme-li každý součet čtverců příslušným počtem stupňů volnosti, dostaneme **průměrný čtverec**, neformálně nazývaný **rozptyl.** Rozptyl mezi výběry (školami) je **vysvětlen** faktem, že jednotlivé výběry pocházejí z různých populací (školy mají různou úroveň matematických dovedností u absolventů). Reziduální rozptyl v jednotlivých výběrech **není vysvětlen**, neboť jde o rozptyl způsobeny náhodnými vlivy.

F-poměr je pak někdy popisován jako poměr vysvětleného a nevysvětleného rozptylu:

Vysvětlený rozptyl (MS_B): $MS_B = \frac{SS_B}{k-1}$,

Nevysvětlený rozptyl (MS_W): $MS_W = \frac{SS_W}{N-k}$,

F-poměr: $F-ratio = \frac{MS_B}{MS_W} = \frac{n \cdot S_B^2}{S_W^2}$

Z toho lze usuzovat na možnosti zesílení jednofaktorové ANOVY. Předpokládejme například, že nevysvětlený rozptyl byl z velké části způsoben rozdílnou úrovni SŠ v jednotlivých městech. Pokud bychom dokázali odstranit rozdíly v úrovni škol v jednotlivých městech, pak by se nevysvětlený rozptyl snížil a tím by se zvýšila hodnota F-poměru. Tzn., měli bychom silnější argument pro zamítnutí H_0.

Schopnost rozpoznat vliv jednoho znaku (školy) lze tudíž posílit zavedením dalšího znaku (města) pro vysvětlení části nevysvětleného rozptylu. Tímto problémem se zabývá dvoufaktorová ANOVA (ANOVA při dvojném třídění). Dvoufaktorovou ANOVOU se zabýváme ve Statistice II.

13.2 Jednofaktorová ANOVA – obecně (pro nestejné rozsahy výběrů)

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

---

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

13.2.1 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.

13.2.2 Zobecněné definiční vztahy

Definujme **totální součet čtverců** (**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:** $SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2$

přičemž S_i^2 je výběrový rozptyl *i*-tého náhodného výběru (i-té třídy): $S_i^2 = \frac{\sum_{i=1}^{n}\left(X_{ij} - \overline{X_i}\right)^2}{n_i - 1}$,

$\overline{X_i}$ je průměr i-tého náhodného výběru (i-té třídy): $\overline{X_i} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$

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

kde **celkový průměr** $\overline{X}$: $\overline{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} n_i \cdot \overline{X_i}}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} n_i \cdot \overline{X_i}}{N}$

(všimněte si že celkový průměr již není prostým průměrem $\overline{X_i}$, ale je váženou formou průměru s váhami n_i)

Zavedeme následující **průměrné součty čtverců** (neformálně rozptyly):

Vnitřní průměrný součet čtverců (vnitřní výběrový rozptyl): $MS_W = \frac{SS_W}{N-k}$

Mezitřídní průměrný součet čtverců (mezitřídní výběrový rozptyl): $MS_B = \frac{SS_B}{k-1}$

Vlastnosti těchto výběrových rozptylů:

- Vnitřní výběrový rozptyl je nestranným odhadem rozptylu, nezávisle na H_0.

$$E[MS_W] = \frac{1}{N-k} E\left[\sum_{i=1}^{k}(n_i - 1)S_i^2\right] = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)E(S_i^2) = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\cdot\sigma^2$$
$$= \frac{\sigma^2}{N-k}(N-k) = \sigma^2$$

- Mezitřídní výběrový rozptyl je nestranným odhadem rozptylu právě když platí H_0.
  $E[MS_W] = \sigma^2 \Leftrightarrow$ když platí H_0

  Mohli bychom dokázat, že $E[MS_B] = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i (E\overline{X} - E\overline{X}_i)^2$,

  z čehož bezprostředně vyplývá následující ekvivalence: $ES_B^2 = \sigma^2 \Leftrightarrow$ když platí H_0

Položíme $F = \frac{MS_B}{MS_W}$

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

Abychom mohli F-poměr v dalším průběhu testu použít jako testovou statistiku (a tím i nulové rozdělení), musíme znát její statistické chování, tedy její rozdělení pravděpodobnosti.

F-poměr má **Fisher-Snedecorovo rozdělení** s počtem stupňů volnosti

pro čitatele **(k-1)** a pro jmenovatele **(N-k)**

Průvodce studiem:

Zájemcům nyní dokážeme, že F-poměr má Fisher-Snedecorovo rozdělení.

Víme, že $\frac{MS_W}{\sigma^2}\cdot(N-k)=\sum_{i=1}^{k}\frac{(n_i-1)S_i^2}{\sigma^2}\to\chi^2_{N-k}$ *, protože* $\frac{(n_i-1)S_i^2}{\sigma^2}\to\chi^2_{n_i-1}$,

dále je známo, že součet náhodných veličin, které mají rozdělení $\chi^2_{n_i-1}$*, je opět náhodnou veličinou stejného typu, s počtem stupňů volnosti daným součtem stupňů volnosti sčítaných veličin.*

Podobnou úvahou lze prokázat, že pokud platí H_0*, pak:*

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

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

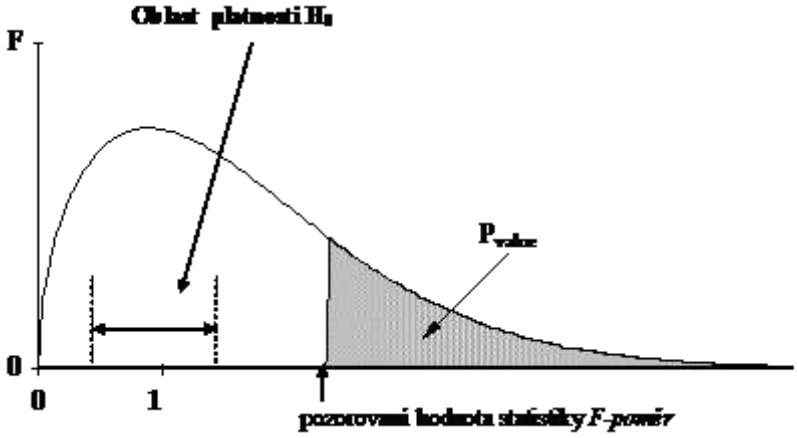
$$\frac{\frac{\frac{MS_B}{\sigma^2}\cdot(k-1)}{k-1}}{\frac{\frac{MS_W}{\sigma^2}\cdot(N-k)}{N-k}}=\frac{MS_B}{MS_W}\to F_{k-1,N-k}$$

Výklad:

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.

Z definice p-value (viz. obrázek) je zřejmé, že:

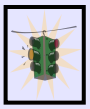
$$p-value = 1 - F(F-ratio)$$



13.2.3 Tabulka ANOVA

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

Zdroj proměnlivosti	Součet čtverců	Stupně volnosti	Průměrný čtverec	Testová stat. F-poměr	P-value
Mezitřídní (faktor)	$SS_B = \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2$	$k-1$	$MS_B = \frac{SS_B}{k-1}$		
Vnitřní (reziduální)	$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2$	$N-k$	$MS_W = \frac{SS_W}{N-k}$	$F-ratio = \frac{MS_B}{MS_W}$	$1-F(F-ratio)$
Celkový	$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2$	$N-1$			



Řešený příklad:

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 vnitřní výběrový výběrový rozptyl, normální a velký rozptyl velký, F-poměr je malý. Datové soubory tak ilustrují tři případy: malý vnitřní.

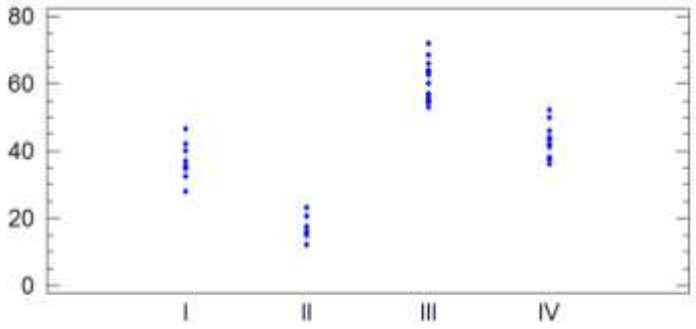
Datový soubor 1: **Malý vnitřní výběrový rozptyl**

	Výběr			
	I	II	III	IV
	42	17,5	68,5	38
	34,5	12	72	44
	32,5	16	53	52
	40	15	64	50
	46,5	20,5	57	43,5
	28	23	56	41
	37	15	54,5	42
	35,5		62,5	46
			63,5	37,5
			60	36
			66	
			55	
Rozsah výběru n_i	8	7	12	10
Průměry $\overline{X_i}$	37,0	17,0	61,0	43,0
Výběrové rozptyly S_i^2	33,4	13,8	36,7	27,7

H_0: $\mu_I = \mu_{II} = \mu_{III} = \mu_{IV}$

H_A: neplatí H_0

	Count	Average	Variance
I	8	37,0	33,4286
II	7	17,0	13,75
III	12	61,0	36,7273
IV	10	43,0	27,7222
Total	37	42,6216	274,242



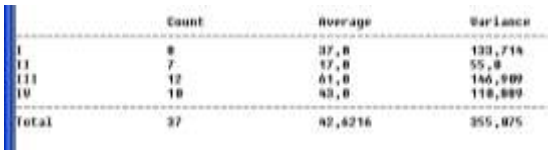
Analysis of Variance

Zdroj	Součet čtverců	Df	Průměrný čtverec	F-poměr	P-Value
Faktor (mezitřídní)	8902,7	3	2967,57	100,96	0,0000
Rezidua (vnitřní)	970,0	33	29,3939		
Celkový	9872,7	36			

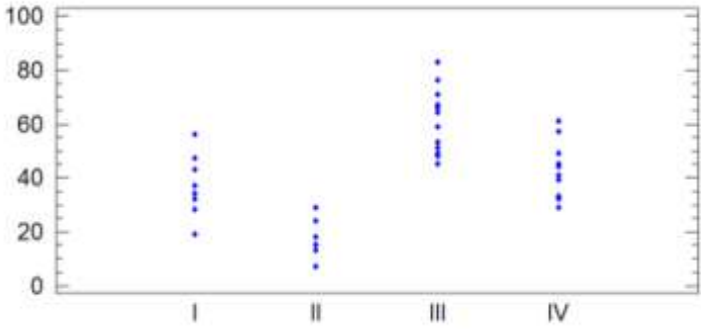
Rozhodnutí: Zamítáme nulovou hypotézu, tzn. dané 4 výběry nepocházejí z jedné populace, jejich střední hodnoty se statisticky významně liší. (Rozptyl mezi třídami je podstatně větší než rozptyl uvnitř tříd.)

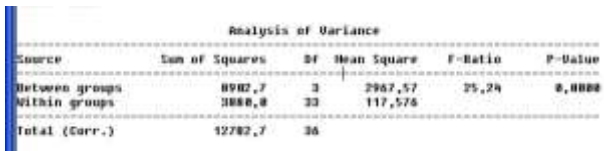
Datový soubor 2: **Normální vnitřní výběrový rozptyl**

	Výběr			
	I	II	III	IV
	47	18	76	33
	32	7	83	45
	28	15	45	61
	43	13	67	57
	56	24	53	44
	19	29	51	39
	37	13	48	41
	34		64	49
			66	32
			59	29
			71	
			49	
Rozsah výběru n_i	8	7	12	10
Průměry $\overline{X_i}$	37,0	17,0	61,0	43,0
Výběrové rozptyly S_i^2	133,7	55,0	146,9	110,9



	Count	Average	Variance
I	8	37,0	133,714
II	7	17,0	55,0
III	12	61,0	146,909
IV	10	43,0	110,889
Total	37	42,4216	355,075





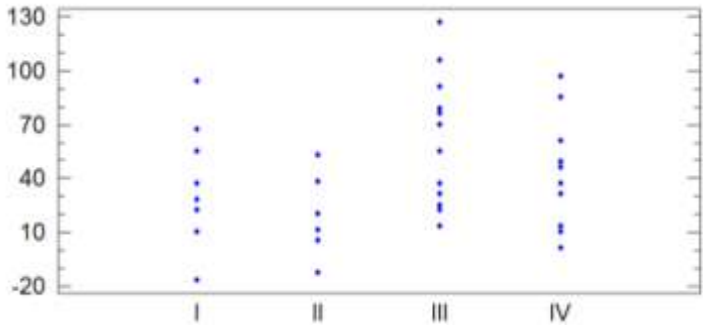
Analysis of Variance

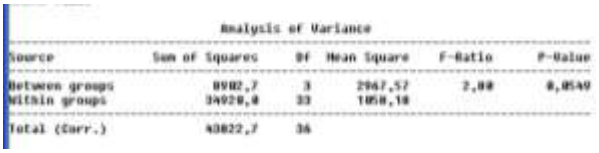
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	8902,7	3	2967,57	25,24	0,0000
Within groups	3880,0	33	117,576		
Total (Corr.)	12782,7	36			

Rozhodnutí: Zamítáme nulovou hypotézu, tzn. dané 4 výběry nepocházejí z jedné populace, jejich střední hodnoty se statisticky významně liší. (Rozptyl mezi třídami je podstatně větší než rozptyl uvnitř tříd.)

Datový soubor 3: **Velký vnitřní výběrový rozptyl**

	Výběr			
	I	II	III	IV
	67	20	106	13
	22	-13	127	49
	10	11	13	97
	55	5	79	85
	94	38	37	46
	-17	53	31	31
	37	5	22	37
	28		70	61
			76	10
			55	1
			91	
			25	
Rozsah výběru n_i	8	7	12	10
Průměry $\overline{X_i}$	37,0	17,0	61,0	43,0
Výběrové rozptyly S_i^2	1203,4	495,0	1322,2	998,0





Rozhodnutí: Nezamítáme nulovou hypotézu, tzn. dané 4 výběry pocházejí z jedné populace, jejich střední hodnoty se statisticky významně neliší. (Rozptyl mezi třídami je srovnatelný s rozptylem uvnitř tříd.)

Výklad:

13.3 Post Hoc analýza (vícenásobné porovnávání)

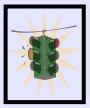
Předchozí příklad poukázal na to, že velký F-poměr indikuje existenci významných změn mezi populačními výběrovými průměry. Naše analýza by ale byla nekompletní, pokud bychom neidentifikovali, které z populací signalizují významnou odchylku průměru. Tento další proces se nazývá **post hoc** analýza a spočívá v porovnávání 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 $(LSD)_{i,j}$**:

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

Snadno lze zdůvodnit, že tato statistika má Studentovo rozdělení s (N-k) stupni volnosti.



Řešený příklad:

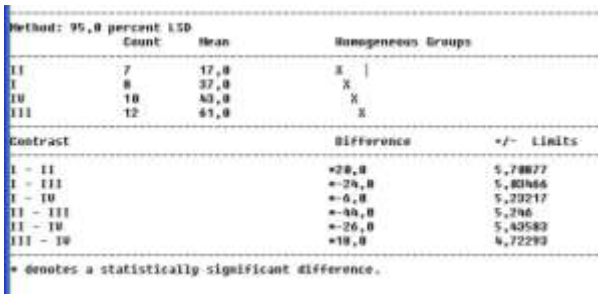
LSD metodu ilustrujeme pro tři předchozí příklady:

Datový soubor 1: **Malý vnitřní výběrový rozptyl**

Zamítli jsme nulovou hypotézu, proto provedeme post hoc analýzu. Vypočteme statistiky $(LSD)_{i,j}$ pro všechny uvažované dvojice daných čtyř populací a hodnoty zaznamenáme do následující tabulky:

	I	II	III	IV
I	0	-7,128	9,698	2,333
II	7,128	0	17,064	9,731
III	-9.698	-17,064	0	-7,754
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.



Method: 95,0 percent LSD

	Count	Mean	Homogeneous Groups
II	7	17,0	X
I	8	37,0	X
IV	10	43,0	X
III	12	61,0	X

Contrast	Difference	+/- Limits
I - II	*20,0	5,78877
I - III	*-24,0	5,80466
I - IV	*-6,0	5,23217
II - III	*-44,0	5,296
II - IV	*-26,0	5,43583
III - IV	*18,0	4,72293

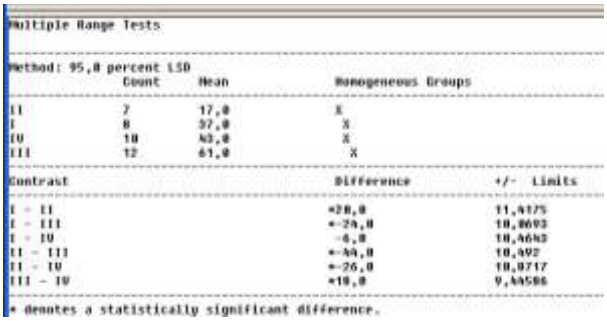
* denotes a statistically significant difference.

Statistický software většinou seskupí výběry, které by mohly pocházet z jedné populace (mají stejné střední hodnoty). Například Statgraphics provádí označení pomocí křížků. Jsou-li v textovém výstupu v části Homogenous Groups (Homogenní výběry) křížky pro příslušné výběry pod sebou, znamená to, že výběry mohou pocházet ze stejné populace.

Datový soubor 2: **Normální vnitřní výběrový rozptyl**

	I	II	III	IV
I	0	-3,564	4,849	1,167
II	3,564	0	8,532	4,8656
III	-4,849	-8,532	0	-3,877
IV	-1,167	-4,866	3,877	0

V tomto případě, ačkoliv 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.



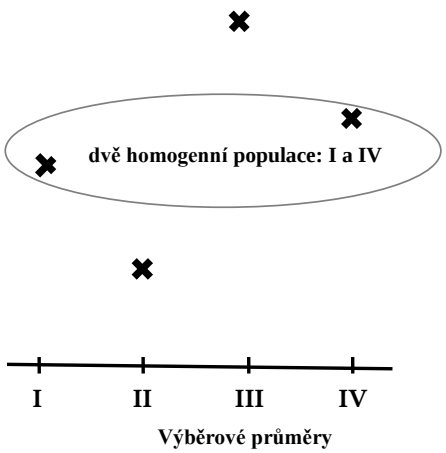
Multiple Range Tests

Method: 95,0 percent LSD

	Count	Mean	Homogeneous Groups
II	7	17,0	X
I	8	37,0	X
IV	10	43,0	X
III	12	61,0	X

Contrast	Difference	+/- Limits
I - II	*20,0	11,4175
I - III	*-24,0	10,8693
I - IV	-6,0	10,4643
II - III	*-44,0	10,492
II - IV	*-26,0	10,0717
III - IV	*18,0	9,44586

* denotes a statistically significant difference.

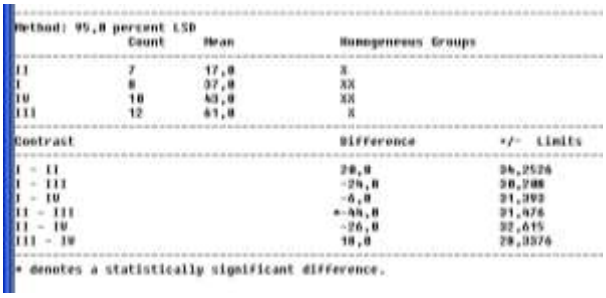


Datový soubor 3: **Velký vnitřní výběrový rozptyl**

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:

	I	II	III	IV
I	0	-1,188	1,616	0,389
II	1,188	0	2,844	1,622
III	-1,616	-2,844	0	-1,292
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**.



Výklad:

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 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 průměrů.

13.4 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.

Kruskal-Wallisův test je tedy neparametrickou obdobou jednofaktorové ANOVY. Na rozdíl od parametrického testu nepředpokládá normalitu rozdělení, jeho nevýhodou je menší citlivost. Tak jako je ANOVA vícevýběrovým testem středních hodnot, Kruskal-Wallisův test je **vícevýběrovým testem mediánů**.

Nechť tyto **náhodné výběry pochází ze spojitých rozdělení stejného typu a stejných rozptylů** (homoskedasticita):

$(X_{11}, X_{12}, ..., X_{1n_1})$
$(X_{21}, X_{22}, ..., X_{2n_2})$
...
$(X_{k1}, X_{k2}, ..., X_{kn_k})$

kde n_i je rozsah jednotlivých výběrů.

Testujeme hypotézu **H_0:** $x_{0,5_I} = x_{0,5_{II}} = \ldots = x_{0,5_{IV}}$
Oproti alternativě **H_A:** neplatí H_0

Všechny veličiny X_{ij} dohromady tvoří sdružený náhodný výběr o rozsahu $N = \sum_{i=1}^{k} n_i$. Z tohoto výběru vytvoříme uspořádaný výběr (rostoucí posloupnost) a určí se **pořadí R_{ij}** každé veličiny X_{ij}. Tato pořadí uspořádáme do tabulky a určíme tzv. **součty pořadí pro jednotlivé výběry** (T_i).

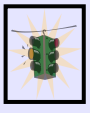
$$T_i = \sum_{j=1}^{n_i} R_{ij}$$

Výběr	Pořadí veličin v uspořádaném sdruženém náhodném výběru				Součty pořadí
1	R_{11}	R_{12}	...	R_{1n1}	T_1
2	R_{21}	R_{22}	...	R_{2n2}	T_2
⋮	⋮	⋮	⋮	⋮	⋮
k	R_{k1}	R_{k2}	...	R_{knk}	T_k

Celkový součet všech pořadí: $T = \sum_{i=1}^{k} T_i = \frac{N \cdot (N+1)}{2}$

Testová statistika: $Q = \frac{12}{N \cdot (N+1)} \cdot \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3 \cdot (N+1) \rightarrow \chi^2_{k-1}$

P-value: $p - value = 1 - F(Q)$



Řešený příklad:

Proveďte Kruskal-Wallisův test pro výše uvedený datový soubor 3.

Řešení:

Výběr			
I	II	III	IV
67	20	106	13
22	-13	127	49
10	11	13	97
55	5	79	85
94	38	37	46
-17	53	31	31
37	5	22	37
28		70	61
		76	10
		55	1
		91	
		25	

Pořadí ve sdruženém výběru:

	Výběr			
	I	II	III	IV
	28	11	36	9,5
	12,5	2	37	23
	6,5	8	9,5	35
	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 n_i	8	7	12	10
Součty pořadí T_i	145	75	293	193,5
T_i^2	20022,3	5625,0	85849,0	37442,3
T_i^2/n_i	2502,8	803,6	7154,1	3744,2

$$Q = \frac{12}{N\cdot(N+1)}\cdot\sum_{i=1}^{k}\frac{T_i^2}{n_i} - 3\cdot(N+1) =$$

$$= \frac{12}{37\cdot(37+1)}\cdot(2502{,}8+803{,}6+7154{,}1+3744{,}2) - 3\cdot(37+1) = 7{,}24$$

$$Q \to \chi^2_{4-1}, \qquad p-value = 1-F(Q) = 0{,}0645$$

P-value pro tuto Q testovou statistiku je o něco větší, než dává F-poměr (0,0549), ale závěry jsou v obou případech stejné. Nulová hypotéza není zamítnuta.

13.5 Postup jednofaktorové analýzy rozptylu (ANOVA)

Na závěr si shrneme zjištěné poznatky.

Vstupem pro analýzu rozptylu je tabulka obsahující pro jednotlivé sloupce (třídy, resp. úrovně sledovaného faktoru) vždy n_i pozorování X_{ij} ((i=1, …, k), kde k je počet tříd; j=1, …, n_i).

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

Postup obsahuje následující kroky:

1. **Příprava dat:** Už přípravou dat lze zajistit větší věrohodnost dosažených výsledků.
   a) ***Velikost výběru*** je počet plných řádků $\left(\min_i \{n_i\}\right)$. ANOVA byla původně navržena pro stejnou četnost v jednotlivých výběrech. V praxi bývá tento předpoklad málokdy splněn – platí však, že čím těsněji je toto pravidlo splněno, tím věrohodnější jsou výsledky. Lze analyzovat i malé výběry (n_i=4, 5). Máme-li však testovat všechny výběrové předpoklady, je třeba mít alespoň 30 hodnot ve výběru $\left(n_i \geq 30\right)$.
   b) ***Chybějící hodnoty*** mohou způsobit vychýlení výsledků.
   c) ***Odlehlé hodnoty*** obecně způsobují nefunkčnost F-testu. Je třeba analyzovat data metodami exploratorní analýzy (EDA) – pro identifikaci odlehlých pozorování použít například krabicový graf. Pokud se odlehlá pozorování vyskytují v datech pouze jednou, je třeba je odstranit. Jestliže je v datech ponecháme, dáme přednost neparametrickému testu (Kruskal-Wallisův test), F-test by mohl selhat.

2. **Ověření výběrových předpokladů:** Nestačí se soustředit na výsledky uvedené v tabulce ANOVA. Je třeba pečlivě ověřit splnění základních předpokladů o výběru. Mnohdy data nemají ve všech výběrech normální rozdělení a je třeba použít vhodnou transformaci (mocninnou, logaritmickou). Po transformaci data vykazují normální rozdělení, což přinese větší důvěryhodnost výsledků.
   a) ***Náhodnost:*** Metoda sběru dat by měla zajistit, že porovnáváme prosté náhodné výběry.
   b) ***Normalita:*** V kapitole Testování hypotéz jsme se seznámili s několika testy normality. Jejich síla roste s velikostí výběrů.
   c) ***Homoskedasticita:*** Homoskedasticita (rovnost rozptylů) je předpokladem pro řadu testů. Numericky lze homoskedasticitu ověřit např. pomoci modifikovaného Levenova testu, popř. pomoci dalších testů nabízených vyspělým statistickým softwarem. Pokud rovnost rozptylů není splněna, mluvíme o heteroskedasticitě.

3. **Volba statistických testů významnosti sledovaného faktoru v tabulce ANOVA**: Naučili jsme se zpracovávat pouze výběry u nichž je splněna:
   a) ***Normalita a homoskedasticita*** – užijeme F-test.
   b) ***Nenormalita a homoskedasticita*** – užijeme Kruskal-Wallisův test
   Pokud homoskedasticita splněna není, pokusíme se pomocí vhodné transformace o stabilizaci rozptylu.
   c) ***Normalita a heteroskedasticita*** – nelze použít ani F-test, ani Kruskal-Wallisův test, neboť homoskedasticita je předpokladem pro oba tyto testy. Pokusíme se rozptyl stabilizovat pomocí vhodné transformace (mocninné, logaritmické). Pokud dojde ke

stabilizaci rozptylu, použijeme F-test na transformovaná data. Pokud se nám rozptyl stabilizovat nepodaří, nelze analýzu rozptylu provést (výsledky nejsou důvěryhodné).

d) ***Nenormalita a heteroskedasticita*** – opět nelze použít ani F-test, ani Kruskal-Wallisův test, neboť homoskedasticita je předpokladem pro oba tyto testy. Pokusíme se rozptyl stabilizovat pomocí vhodné transformace (mocninné, logaritmické). Pokud dojde ke stabilizaci rozptylu, došlo-li zároveň k normalizaci dat, použijeme F-test na transformovaná data, nedošlo-li k normalizaci dat, použijeme Kruskal-Wallisův test na transformovaná data. Pokud se nám rozptyl stabilizovat nepodaří, nelze analýzu rozptylu provést (výsledky nejsou důvěryhodné).

4. **Post hoc analýza (vícenásobné porovnávání)**: Pokud při analýze rozptylu došlo k zamítnutí nulové hypotézy, pokoušíme se pomocí vícenásobného porovnávání nalézt homogenní (srovnatelné) populace. Vícenásobné porovnávání předpokládá normalitu a homoskedasticitu výběrů. Není-li splněn předpoklad normality, je třeba užít Kruskal-Wallisův test vícenásobného porovnávání.

Shrnutí:

Rozšířením dvouvýběrových testů pro střední hodnoty je **analýza rozptylu** neboli **ANOVA,** která umožňuje srovnávat několik středních hodnot nezávislých náhodných výběrů. Analýza rozptylu ve své parametrické podobě **předpokládá normalitu rozdělní a** tzv. **homoskedasticitu** (identické rozptyly).

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, získané 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.** Post hoc analýza se provádí pouze v případě zamítnutí H_0. Použijeme-li ji v případě, kdy H_0 nezamítneme, můžeme dostat **falešné výsledky**.

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, se používá **Kruskal - Wallisův** pořadový test**.**

Otázky

1. Co je to ANOVA?
2. Popište konstrukci a stochastické chování statistiky F-poměr.
3. Co je to vnitřní a mezitřídní výběrový rozptyl ?
4. Jaký je obvyklý výstup z analýzy rozptylu ?
5. Co je to post hoc analýza a LSD-statistika ?
6. Co je to Kruskal-Wallisův test, kdy se používá?

Úlohy k řešení

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
10,9	8,9	11,2
9,8	10,3	9,7
6,4	7,5	15,8
4,3	6,9	8,9
7,5	14,1	12,2
12,3	9,3	17,5
5,1	12,5	10,1

   Rozhodněte, zda vzdělání má vliv na příjem.

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, nezapomeňte ověřit předpoklady testu}

Počet členů domácnosti	**Výdaje na jednoho člena domácnosti (v Kč)**				
	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	

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			

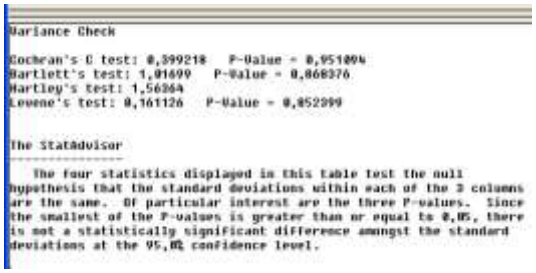
   Pokuste se prokázat existenci rozdílů v nákladech mezi jednotlivými typy bytů.

(Použijte vhodný programový balík, nezapomeňte ověřit předpoklady testu)

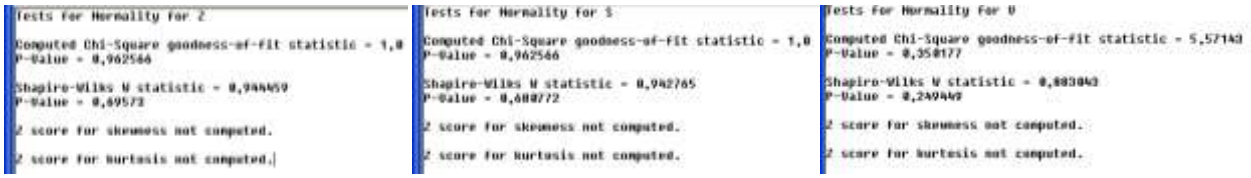
Řešení:

ad1)
Ověření předpokladů:

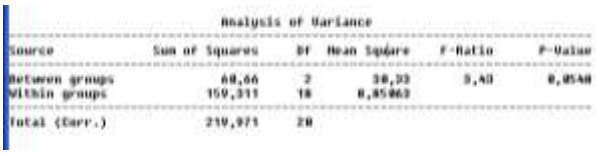
Homoskedasticita:



Normalita:



Normalita i homoskedasticita potvrzena, tzn. můžeme použít F-test.

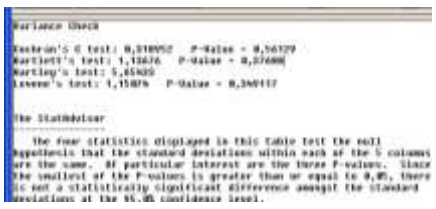


H_0: $\mu_Z = \mu_S = \mu_V$
H_A: neplatí H_0

Rozhodnutí: Nezamítáme H_0, tzn. na základě předložených dat musíme říci, že s 95% ní spolehlivosti vzdělání nemá vliv na výši platu.

ad2)
Ověření předpokladů:

Homoskedasticita:



Normalita:

Pro 4. výběr byla normalita zamítnuta, homoskedasticita byla potvrzena, tzn. můžeme použít Kruskal-Wallisův test.

H_0: $x_{0,5_I} = x_{0,5_{II}} = x_{0,5_{III}} = x_{0,5_{IV}} = x_{0,5_V}$
H_A: neplatí H_0

Kruskal-Wallis Test

	Sample Size	Average Rank
I	5	38,0
II	8	23,75
III	10	23,7
IV	10	15,25
V	7	7,21429

Test statistic = 23,6316 P-Value = 0,0000946781

The StatAdvisor

The Kruskal-Wallis test tests the null hypothesis that the medians within each of the 5 columns is the same. The data from all the columns is first combined and ranked from smallest to largest. The average rank is then computed for the data in each column. Since the P-value is less than 0,05, there is a statistically significant difference amongst the medians at the 95,0% confidence level. To determine which medians are significantly different from which others, select Box-and-Whisker Plot from the list of Graphical Options and select the median notch option.

Rozhodnutí: Zamítáme H_0, tzn. na základě předložených dat můžeme říci, že s 95% ní spolehlivosti má počet členů domácnosti vliv na velikost průměrných (na osobu) měsíčních výdajů domácnosti za potraviny.

Zamítli jsme H_0, proto provedeme post hoc analýzu (použili jsme Tukey HSD test):

Multiple Range Tests

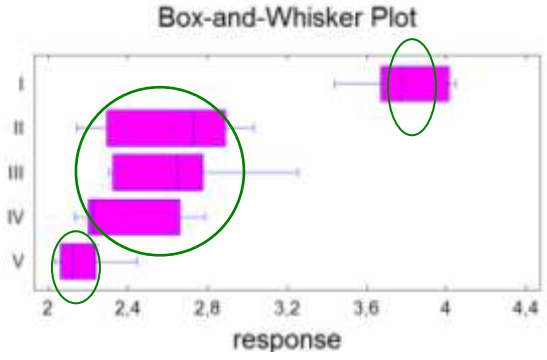
Method: 95,0 percent Tukey HSD

	Count	Mean	Homogeneous Groups
V	7	2,16286	X
IV	10	2,3792	XX
II	8	2,62437	X
III	10	2,6451	X
I	5	3,7892	X

Contrast	Difference	+/- Limits
I - II	*1,16483	0,447325
I - III	*1,1441	0,429776
I - IV	*1,41	0,429776
I - V	*1,62634	0,45945
II - III	-0,020725	0,372197
II - IV	0,245175	0,372197
II - V	*0,461518	0,4061
III - IV	0,2659	0,358911
III - V	*0,482243	0,386685
IV - V	0,216343	0,386685

* denotes a statistically significant difference.

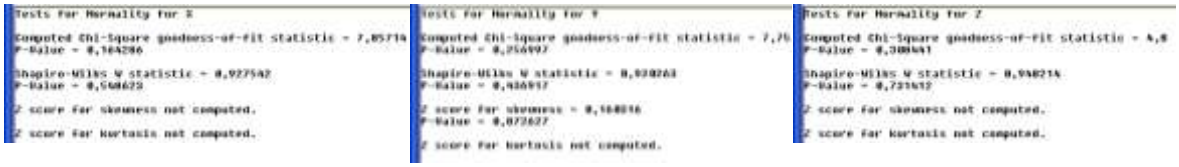
Tukeyho test nám ukázal, že data můžeme považovat za výběry ze tří populací. Do jedné skupiny můžeme zařadit jednočlenné domácnosti, v nichž jsou průměrné náklady na potraviny nejvyšší, do druhé skupiny zařadíme např. dvou až čtyřčlenné domácnosti a jako třetí skupinu budeme uvažovat pětičlenné domácnosti, jejichž průměrné měsíční výdaje za potraviny jsou nejnižší. (viz. obr.)



ad3)
Ověření předpokladů:

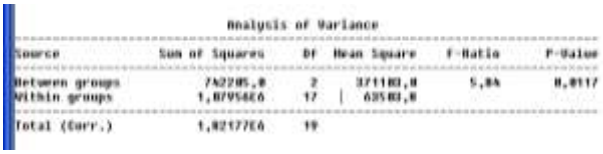
Homoskedasticita:

Normalita:



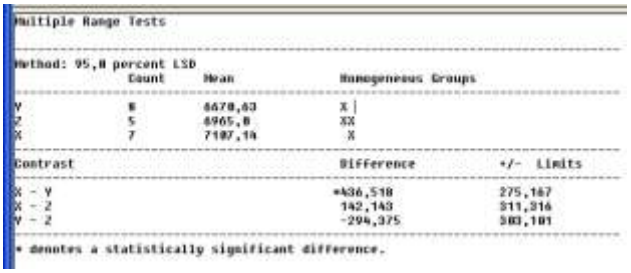
Normalita i homoskedasticita potvrzena, tzn. můžeme použít F-test.

H_0: $\mu_X = \mu_Y = \mu_Z$
H_A: neplatí H_0



Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	742285,8	2	371143,8	5,84	0,0117
Within groups	1,079566E6	17	63503,8		
Total (Corr.)	1,82177E6	19			

Rozhodnutí: Zamítáme H_0, tzn. na základě předložených dat můžeme říci, že s 95% ní spolehlivosti má tyb bytů vliv na náklady na $1m^2$.
Zamítli jsme H_0, proto provedeme post hoc analýzu:



Multiple Range Tests

Method: 95,0 percent LSD

	Count	Mean	Homogeneous Groups
Y	8	6670,63	X
Z	5	6965,0	XX
X	7	7107,14	X

Contrast	Difference	+/- Limits
X - Y	*436,518	275,167
X - Z	142,143	311,314
Y - Z	-294,375	303,101

* denotes a statistically significant difference.

Je zřejmé, že příčinou rozdílů jsou rozdíly mezi byty typu X a byty typu Y. Rozdělení může vypadat například takto: Samostatnou skupinou jsou byty typu X, jejichž náklady na $1m^2$ jsou významně (správněji statisticky významně) vyšší než náklady na $1m^2$ bytu typu Y, resp. Z. Náklady na na $1m^2$ bytu typu Y a bytu typu Z můžeme považovat za totožné. (viz. obr.)