

OBSAH

Charakteristika a význam analýzy rozptylu

Vzorové údaje 1.6

Praktické použitie programu SAS (SAS Enterprise Guide)

Príklad 1.1 (SAS) - jednofaktorová analýza rozptylu

Príklad 1.2 (SAS) - viacfaktorová analýza rozptylu

Praktické použitie programu EXCEL (Microsoft 365)

Príklad 1.1 (Excel)

Príklad 1.2 (Excel)

Praktické použitie programu R (R Studio)

Príklad 1.1 (Program R)

Príklad 1.2 (Program R)

Ďalšie príklady použitia analýzy rozptylu

Zdroje a zoznam použitej literatúry

Charakteristika a význam analýzy rozptylu (ANOVA - Analysis of Variance)

Analýza rozptylu je parametrická štatistická metóda vytvorená na vzájomné porovnávanie znakov viacerých skupín (faktorov, činiteľov), ktorých počet je väčší ako dve. Veľmi zjednodušene sa môže povedať, že je to v skutočnosti akoby T-test pre viac ako dve skupiny výberových údajov. Analýza rozptylu patrí medzi základné metódy štatistickej analýzy údajov. ANOVA nachádza veľmi široké uplatnenie prakticky vo všetkých oblastiach aplikovanej štatistiky.

V oblasti biologického výskumu a analýzy údajov má nezastupiteľné miesto prakticky vo všetkých komplexnejších analýzach chápania samotných biologických problémov a javov. Je neodmysliteľná pri rôznych typoch genetických hodnotení zvierat. Princíp odhadu plemenných hodnôt (odhad genetickej kvality zvierat) je vo svojej podstate založený na využívaní rôznych typov a modifikácií analýzy rozptylu.

Existuje viacero typov analýzy rozptylu. Najjednoduchším typom analýzy rozptylu je jednofaktorová analýza rozptylu (One-Way ANOVA). Analýza rozptylu, ktorá obsahuje viac ako jeden faktor sa nazýva viacfaktorová analýza rozptylu.

Jednofaktorová analýza rozptylu nám umožňuje testovať hypotézy o rovnostiach stredných hodnôt viacerých skupín. Cieľom analýzy rozptylu je zistiť, či rozdiely, ktoré sme vo viacerých skupinách zistili, sú, alebo nie sú štatisticky významné. Ak nie sú štatisticky významné tak sú iba iba výsledkom náhody.

Jednofaktorová je vo svojej podstate najjednoduchšou formou ANOVA, ktorá skúma vzťah medzi intervalovou a nominálnou premennou. Prakticky testuje nulovú hypotézu o zhode stredných hodnôt, pričom predpokladá, že výbery majú rovnaký rozptyl. Nulová hypotéza vyjadruje, že medzi intervalovou a nominálnou premennou nie je vzťah. V prípade, že nulovú hypotézu musíme odmietnuť platí hypotéza jednotková, že medzi intervalovou a nominálnou premennou je štatisticky významný, alebo dokonca vysoko štatisticky významný vzťah.

Ak je výsledkom analýzy rozptylu zamietnutie nulovej hypotézy a záver, že existujú rozdiely medzi priermi skupín, môžeme ďalej použiť viaceré porovnávacie testy a metódy na určenie, ktoré konkrétne páry priemerov skupín sú preukazne odlišné.

Najčastejšie sa využíva metóda LSD (Least Significant Difference, metóda najmenej významného rozdielu), ktorá je ekvivalentná vykonávaniu párových T-testov medzi všetkými skupinami. Ďalšími porovnávacími metódami sú: Tukeyho a Bonferroniho metóda. Tukey a Bonferroni metódy sú vhodnejšie, ak sa analyzuje pomerne veľké množstvo porovnaní.

Analýzy rozptylu, tak ako aj iné štatistické analýzy, je založená na splnení viacerých predpokladov:

1. Skupiny údajov musia byť náhodné a nezávislé.
2. Údaje neobsahujú žiadne mimoriadne extrémne hodnoty.
3. Skupiny údajov majú približne rovnaký rozptyl (homoskedasticita).

Narušenie rovnakého rozptylu nevedí ak sú skupiny vyvážené (rovnako veľké), alebo použijeme Welchov F-test

4. Analyzované pozorovania pochádzajúce v rámci skupín z normálneho rozloženia početnosti (pri použití analýzy rozptylu skúmame v rámci výsledkov normalitu reziduálnej zložky variability).

Narušenie normality nevedí, ak sú skupiny rovnako veľké a majú minimálnu veľkosť 30 pozorovaní, neparametrická možnosť – použitie Kruskal-Wallisovho testu

Splnenie predpokladov analýzy rozptylu sa musí overiť ešte pred samotným použitím metódy.

Zostavenie grafov a základná popisná štatistika údajov nám môže pomôcť overiť predpoklad približne rovnakých rozptylov pre každú skupinu. Pokiaľ chceme exaktne upresniť tento predpoklad musíme by použiť test homogenity rozptylov. Najčastejšie sa používajú Leveneov, Brown-Forsytheov test, alebo Bartlettov test na zistenie homogenity rozptylov.

Testy sú ale pomerne rozdielne a ich výsledky a použitie môže byť pri rovnakých údajoch značne odlišné. Výsledok ich interpretácie môže viesť k úplne opačnému záveru. Bartlettov test je mimoriadne citlivý na dodržanie normality údajov, čo nie je tak prísne definované pre použitie Leveneovho testu. Predpoklad zhodných rozptylov ale nie je úplne nevyhnutný a v skutočnosti existuje modifikácia analýzy rozptylu, ktorá sa môže použiť aj bez dodržania predpokladu rovnakých rozptylov.

Bartlettov test je modifikáciou testu pomeru normálnej teórie. Tento test vypočíta presnú chybovosť typu I, keď je distribúcia údajov normálna.

Brown-Forsytheov test je variáciou Leveneovho testu. Rovnaké rozptyly sa určujú pomocou absolútnych odchýlok od stredových hodnôt skupiny. Hoci je to dobrý test na určenie rozdielov rozptylu, môže byť náročný na zdroje, ak vaše údaje obsahujú niekoľko rôznych veľkých skupín.

Leveneov test sa považuje za štandardný test homogenity rozptylu. Tento test počíta druhú mocninu rezíduí, aby sa určil rovnaký rozptyl.

Ostatné predpoklady analýzy rozptylu, náhodné, nezávislé pozorovania a normálne rozložené početnosti údajov možno tiež overiť zostavením grafov reziduí spolu s odhadnutými hodnotami.

V tabuľke 6.1 uvádzame základné číselné informácie, parametre a hodnoty s ktorými sa stretávame pri analýze rozptylu.

Tab. 6.1 Základná tabuľka analýzy rozptylu

Zdroj premenlivosti	Stupne voľnosti df	Suma štvorcov SS	Priemerný štvorec MS	F test	p hodnota
Model (medzi skupinami)	$m - 1$	SS_M	MS_M	MS_M / MS_E	
Chyba (vnútri skupín)	$n - m$	SS_E	MS_E		
Spolu	$n - 1$	SS_T	MS_T		

Variabilita medzi skupinami (Model) je vyjadrená ako suma štvorcov rozdielov medzi priemerom pre každú skupinu a celkovým priemerom (SS, variabilita ktorá je vysvetlená nezávisle premennou). Variabilita vo vnútri skupín (chyba) je vyjadrená ako suma štvorcov rozdielov medzi každou pozorovanou hodnotou a priemerom pre príslušnú skupinu (variabilita, ktorá je nevysvetlená). Tieto rozdiely sa označujú ako reziduá. Celková variabilita (spolu) je suma štvorcov rozdielov medzi každou pozorovanou hodnotou a celkovým priemerom.

Stupne voľnosti (df) sú dané rozdielom počtu pozorovaní v analyzovanom súbore a počtom obmedzení, ktorý je daný počtom použitých štatistík vypočítaných z výberových údajov. Celkový počet stupňov voľnosti je o jeden menší ako celkový počet údajov vo výberovom súbore ($n - 1$). Počet stupňov voľnosti v rámci modelu je o jeden menší ako počet použitých skupín ($m - 1$). Počet stupňov voľnosti chyby je celkový počet údajov mínus počet skupín ($n - m$).

Hodnota p určuje, na akej hladine významnosti je možné zamietnúť nulovú hypotézu o zhode stredných hodnôt. Porovnáva sa s dopredu stanoveným číslom (najčastejšie s hodnotou 0,05 resp. 0,01). Ak je hodnota menšia, alebo rovná stanovenej hranici tak nulovú hypotézu zamietame.

Vzorové údaje 1.6 (zdroj: Databáza údajov projektu KEGA)

Vzorové údaje predstavuje súbor údajov telesných mier študentov (TM.xlsx), ktorý obsahu nasledovné premenné:

PORADIE	poradové číslo záznamu
POHL	muž - M, žena - Z
VEK	vek v rokoch
ROK	rok merania
HMOT	hmotnosť (kg)
VYSKA	telesná výška (cm)
OBVODH	obvod hrudníka (cm)
OBVODP	obvod pásu (cm)
OBVODZ	obvod zápästia (cm)
KRUZOK	študijný krúžok
BMI	index telesnej hmotnosti (Body Mass Index)
INDEXHP	relatívny index (pomer obvodu hrudníka a obvodu pásu, %)

Ukážka časti súboru údajov (10 záznamov):

PORADIE	POHL	VEK	ROK	HMOT	VYSKA	OBVODH	OBVODP	OBVODZ	KRUZOK	BMI	INDEXHP
1	Z	21	2008	52	160	90	70	15.0	1	20.3	128.6
2	Z	20	2008	54	174	86	68	16.0	1	17.8	126.5
3	Z	22	2008	57	174	83	63	16.0	1	18.8	131.7
4	Z	21	2008	62	168	100	76	15.0	1	22.0	131.6
5	M	21	2008	67	169	94	86	18.0	1	23.5	109.3
6	M	21	2008	72	192	94	89	17.0	1	19.5	105.6
7	M	20	2008	75	176	94	85	19.0	1	24.2	110.6
8	M	21	2008	81	187	100	87	21.0	1	23.2	114.9
9	M	22	2008	89	180	105	104	19.0	1	27.5	101.0
10	M	20	2008	65	186	97	78	18.0	3	18.8	124.4

Praktické použitie programu SAS (SAS Enterprise Guide)**Príklad 1.1 (SAS) - jednofaktorová analýza rozptylu**

Úlohy (Tasks): Analyze - ANOVA - One-Way ANOVA

Overte nulovú hypotézu, že neexistuje štatisticky preukazný rozdiel medzi priemernými hodnotami ukazovateľa BMI (index telesnej hmotnosti) v skupinách študentov podľa rokov zisťovania telesných mier (použite hranicu spoľahlivosti 95 %, $\alpha = 0,05$). Rovnakú hypotézu overte aj v triedení podľa pohlavia študentov.

V tabuľke 6.2 uvádzame početnosť podľa rokov zisťovania telesných mier.

Tab. 6.2 Početnosti skupín podľa rokov

ROK	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2008	62	12.68	62	12.68
2009	92	18.81	154	31.49
2011	66	13.50	220	44.99
2012	135	27.61	355	72.60
2014	134	27.40	489	100.00

V tabuľkách 6.3 a 6.4 uvádzame početnosť podľa rokov zisťovania telesných mier v triedení podľa pohlavia.

Tab. 6.3 Početnosti skupín podľa rokov - pohlavie M

ROK	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2008	29	17.68	29	17.68
2009	48	29.27	77	46.95
2011	22	13.41	99	60.37
2012	36	21.95	135	82.32
2014	29	17.68	164	100.00

Tab. 6.4 Početnosti skupín podľa rokov - pohlavie Z

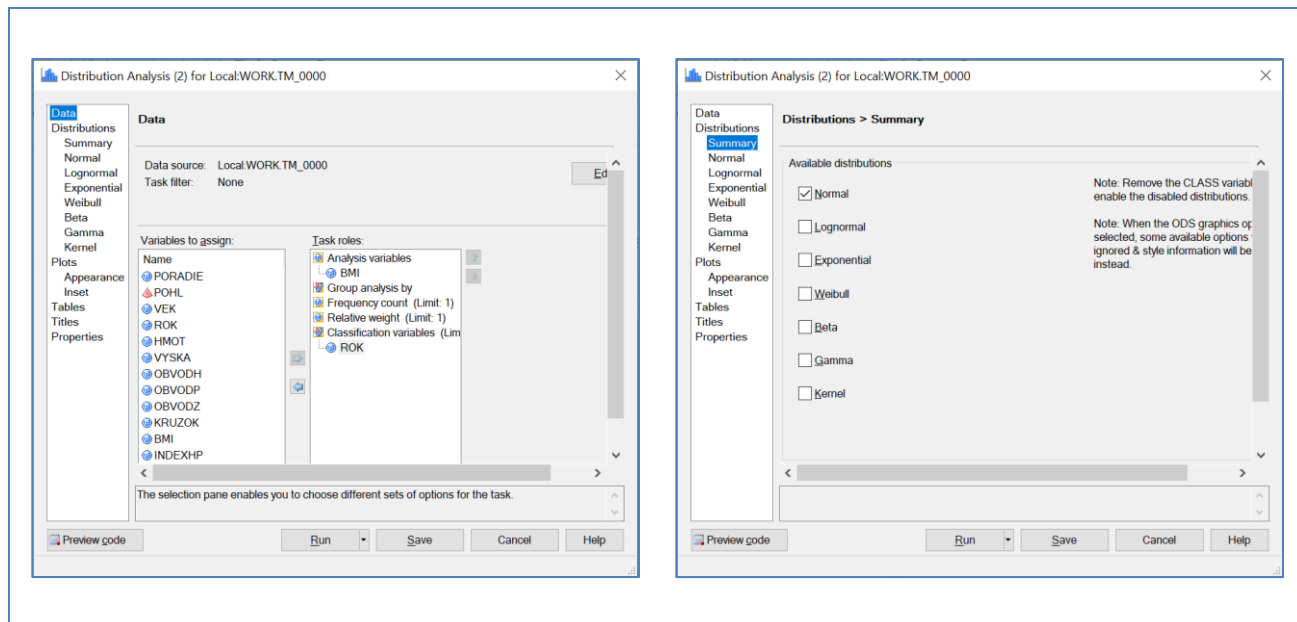
ROK	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2008	33	10.15	33	10.15
2009	44	13.54	77	23.69
2011	44	13.54	121	37.23
2012	99	30.46	220	67.69
2014	105	32.31	325	100.00

Pred samotným použitím jednofaktorovej analýzy rozptylu musíme uskutočniť overenie predpokladu, že hodnotený ukazovateľ má približne normálne rozdelenie počtosti. Na overenie predpokladu použijeme distribučnú analýzu hodnoteného ukazovateľa BMI.

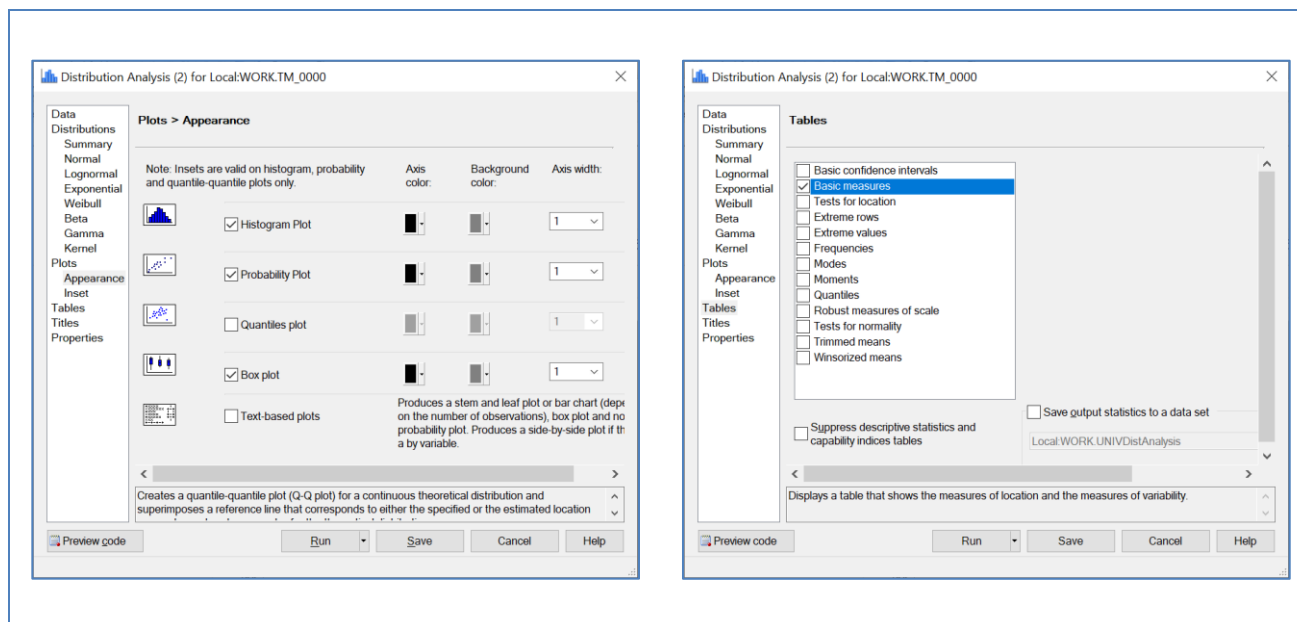
Úlohy (Tasks): Describe - Distribution Analysis

Distribučná analýza úloha má nasledovné základné nastavenia a parametre (tabuľky 6.5a 6.5b):

Tab. 6.5a Nastavenie distribučnej analýzy

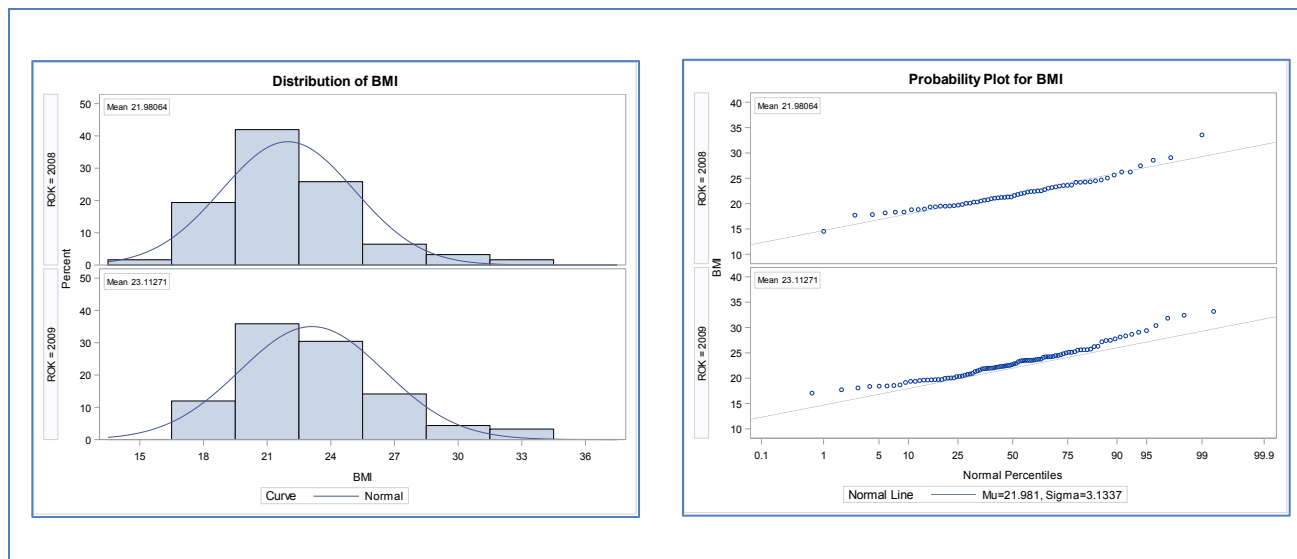


Tab. 6.5b Nastavenie distribučnej analýzy



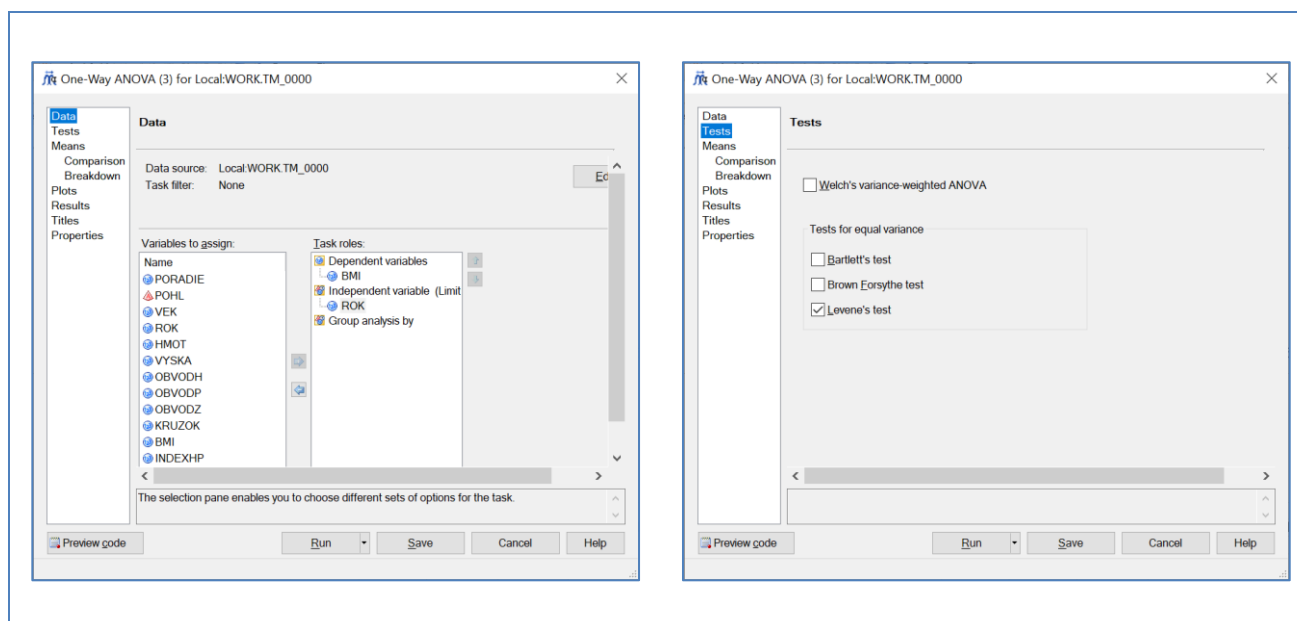
Výsledky distribučnej analýzy (histogramy a grafy pravdepodobnosti percentíl, z dôvodu väčšieho rozsahu uvádzame iba pre roky 2008 a 2009).

Tab. 6.6 Grafické výsledky distribučnej analýzy

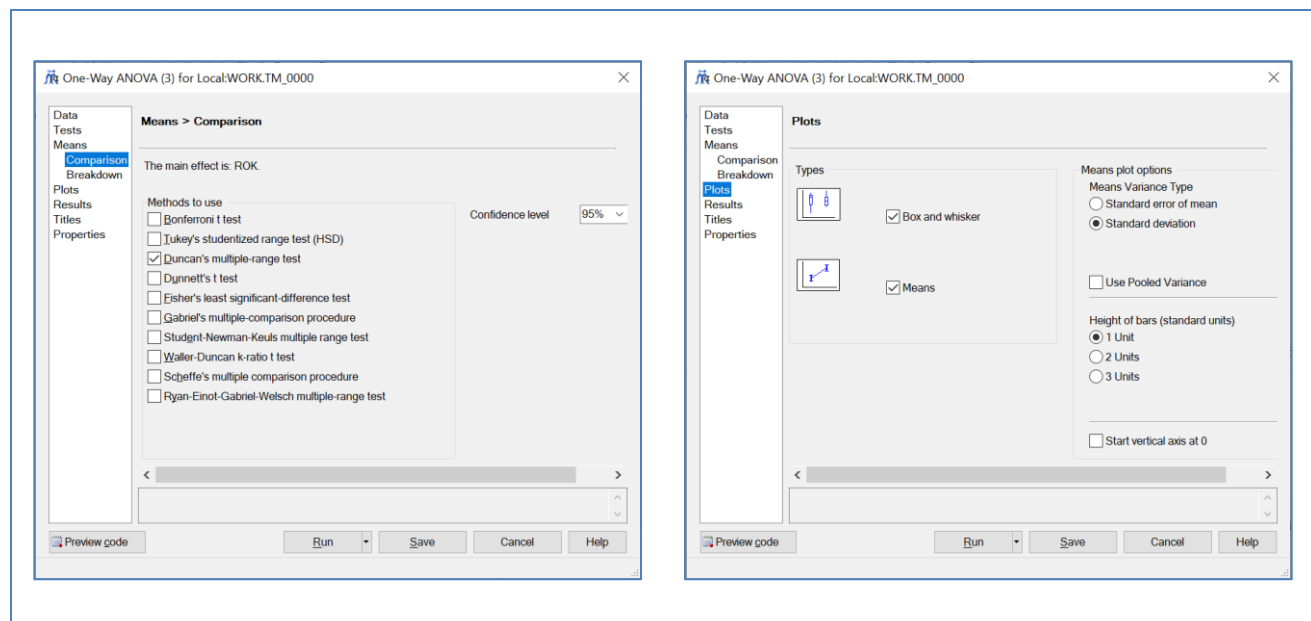


Na základe grafického znázornenia distribučnej analýzy môžeme konštatovať, že hodnotený ukazovateľ má približne normálne rozdelenie a preto môžeme uskutočniť samotnú jednofaktorovú analýzu rozptylu. Potvrdzujú to aj grafy pravdepodobnosti normality percentíl, pretože jednotlivé body ležia v približnej priamke, čo naznačuje, že údaje sú približne normálne rozložené. Grafické posúdenie normality je úplne postačujúce. Číselné testy potvrdenie normality nemusíme v tomto prípade použiť. Jednofaktorová analýza má nasledovné základné nastavenia (tabuľky 6.7 a 6.7b):

Tab. 6.7a Nastavenie jednofaktorovej analýzy rozptylu



Tab. 6.7b Nastavenie jednofaktorovej analýzy rozptylu



Výsledky jednofaktorovej analýzy rozptylu a ich interpretácia

Výsledky jednofaktorovej analýzy rozptylu tvorí informácia o počte a identifikácii jednotlivých skupín analyzovaného faktora, počet použitých pozorovaní a základná tabuľka analýzy rozptylu (tabuľka 6.8).

Tab. 6.8 Výsledky jednofaktorovej analýzy rozptylu

Class Level Information					
Class	Levels	Values			
ROK	5	2008	2009	2011	2012 2014

Number of Observations Read	489
Number of Observations Used	489

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	52.702770	13.175693	1.07	0.3716
Error	484	5970.091949	12.334901		
Corrected Total	488	6022.794719			

R-Square	Coeff Var	Root MSE	BMI Mean
0.008751	15.55434	3.512108	22.57960

Na základe tabuľky analýzy rozptylu môžeme konštatovať, že nulová hypotéza sa potvrdila. Neexistuje štatisticky preukazný rozdiel medzi priemernými hodnotami ukazovateľa BMI v skupinách študentov podľa rokov zisťovania telesných mier ($F\text{-test}=1,07$, $p=0,3716$). Koeficient determinácie dosiahol prakticky nulovú hodnotu, čo znamená, že rok zisťovania telesných mier študentov má minimálny vplyv na variabilitu (rozptyl) hodnoteného znaku.

Výsledok testu homogenity rozptylov jednotlivých analyzovaných skupín uvádzame v tabuľke 6.9).

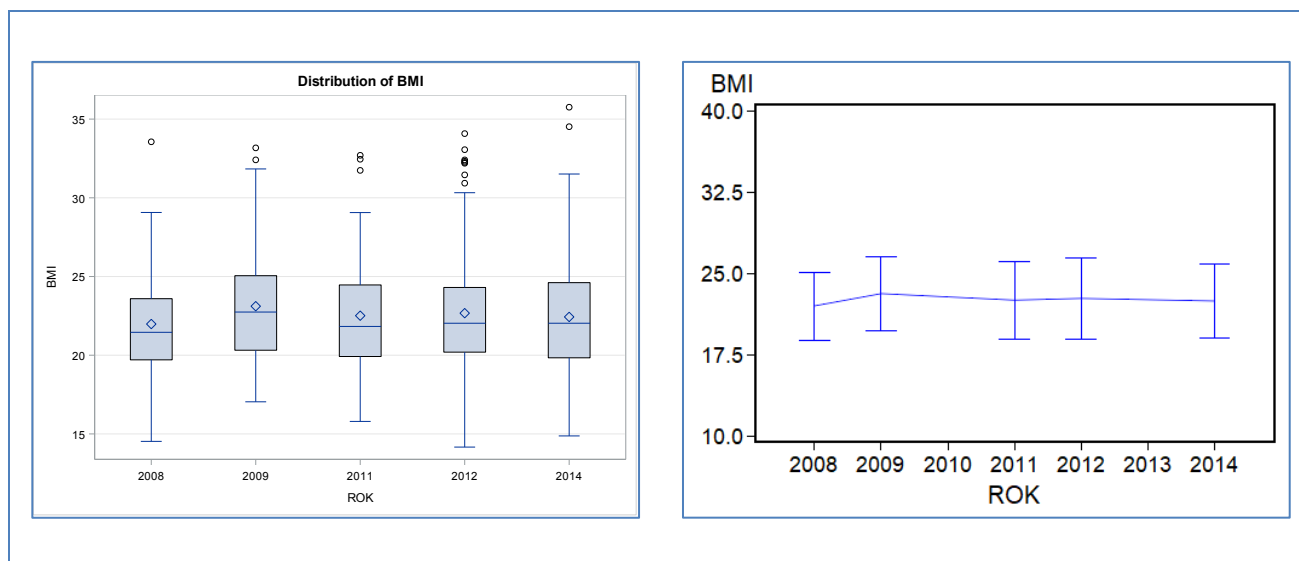
Tab. 6.9 Výsledky testu homogenity rozptylov

Levene's Test for Homogeneity of BMI Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ROK	4	831.7	207.9	0.46	0.7649
Error	484	218629	451.7		

Výsledky Leveneovho testu ($F\text{-test}=0,46$, $p=0,7649$) potvrdili, že platí homogenita rozptylov jednotlivých hodnotených skupín a analýza rozptylu bola z tohto hľadiska použitá oprávnenne.

Priemerné hodnoty a variabilitu ukazovateľa BMI podľa hodnotených skupín uvádzame aj v grafickej podobe.

Tab. 6.10 Grafické výsledky jednofaktorovej analýzy rozptylu



Uvedené grafické výsledky potvrdzujú tiež iba minimálne rozdiely priemerných hodnôt medzi jednotlivými rokmi.

Konkrétne hodnoty rozdielov priemerných hodnôt spolu s viacnásobnými porovnávacími testami a ich intervalmi spoľahlivosti uvádzame v tabuľke. Použili sme Fisherov LSD test (Fisher's least-significant-difference) na hladine 95 %.

Tab. 6.11 Výsledky jednofaktorovej analýzy rozptylu (porovnania)

Comparisons significant at the 0.05 level are indicated by ***.			
ROK Comparison	Difference Between Means	95% Confidence Limits	
2009 - 2012	0.4421	-0.4909	1.3750
2009 - 2011	0.6021	-0.5110	1.7153
2009 - 2014	0.6797	-0.2546	1.6141
2009 - 2008	1.1321	-0.0018	2.2660
2012 - 2009	-0.4421	-1.3750	0.4909
2012 - 2011	0.1601	-0.8764	1.1965
2012 - 2014	0.2376	-0.6039	1.0792
2012 - 2008	0.6900	-0.3687	1.7487
2011 - 2009	-0.6021	-1.7153	0.5110
2011 - 2012	-0.1601	-1.1965	0.8764
2011 - 2014	0.0776	-0.9602	1.1153
2011 - 2008	0.5299	-0.6906	1.7504
2014 - 2009	-0.6797	-1.6141	0.2546
2014 - 2012	-0.2376	-1.0792	0.6039
2014 - 2011	-0.0776	-1.1153	0.9602
2014 - 2008	0.4523	-0.6076	1.5123
2008 - 2009	-1.1321	-2.2660	0.0018
2008 - 2012	-0.6900	-1.7487	0.3687
2008 - 2011	-0.5299	-1.7504	0.6906
2008 - 2014	-0.4523	-1.5123	0.6076

Z výsledkov v tabuľke je zrejmé, že neexistuje žiadna štatistická preukaznosť medzi rozdielmi priemerov jednotlivých porovnávaných rokov.

Rovnaké výsledky (nulová hypotéza) boli potvrdené aj pri použití identickej jednofaktorovej analýzy rozptylu uskutočnenej v triedení podľa pohlavia študentov. V prípade pohlavia M dosiahol F-test hodnotu 1,33 ($p=0,2609$) a pri pohlaví Z F-test dosiahol hodnotu 1,13 ($p=0,3435$). Podrobné výsledky tejto jednofaktorovej analýzy rozptylu uvádzame v prílohe Modulu 1_6.

Príklad 1.2 (SAS) - viacfaktorová analýza rozptylu

Úlohy (Tasks): Analyze - ANOVA – Linear Models

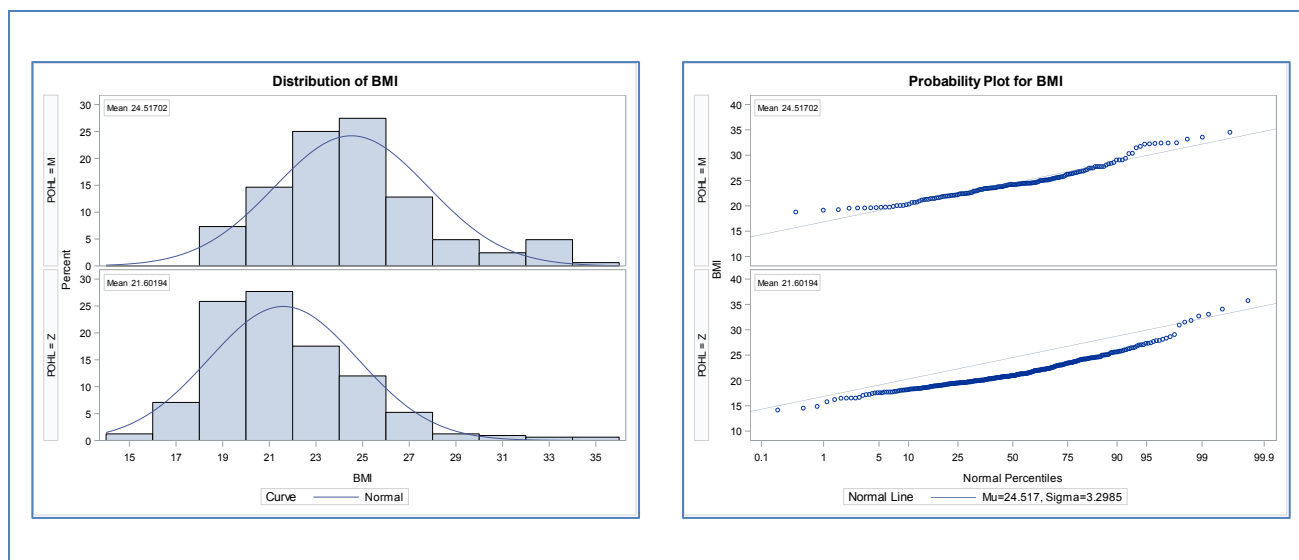
Overte pomocou viacfaktorovej analýzy rozptylu nulovú hypotézu, že neexistujú štatisticky preukazné rozdiely medzi priemernými hodnotami ukazovateľa BMI v skupinách študentov podľa rokov zisťovania telesných mier a podľa pohlavia študentov spolu (použite hranicu spoľahlivosti 95 %, $\alpha = 0,05$). V tabuľke 6.12 uvádzame početnosť záznamov podľa pohlavia študentov.

Tab. 6.12 Početnosti skupín podľa pohlavia

POHL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
M	164	33.54	164	33.54
Z	325	66.46	489	100.00

Výsledky distribučnej analýzy podľa pohlavia (histogramy a grafy pravdepodobnosti percentíl, tabuľka 6.13)

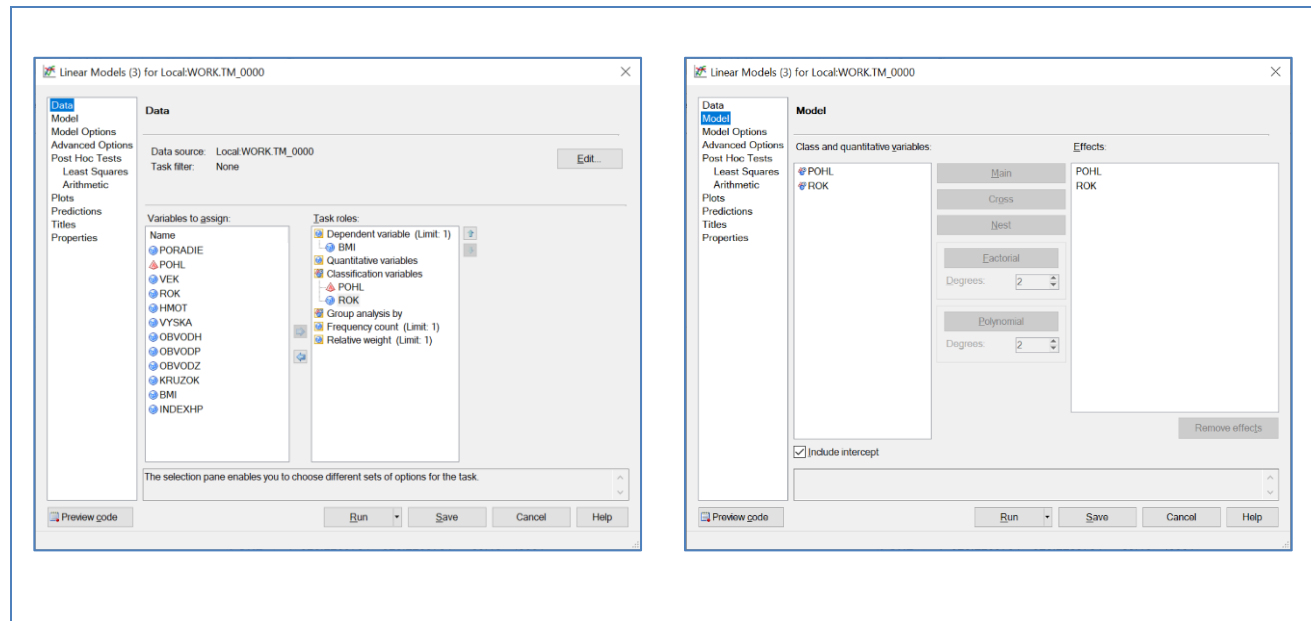
Tab. 6.13 Grafické výsledky distribučnej analýzy



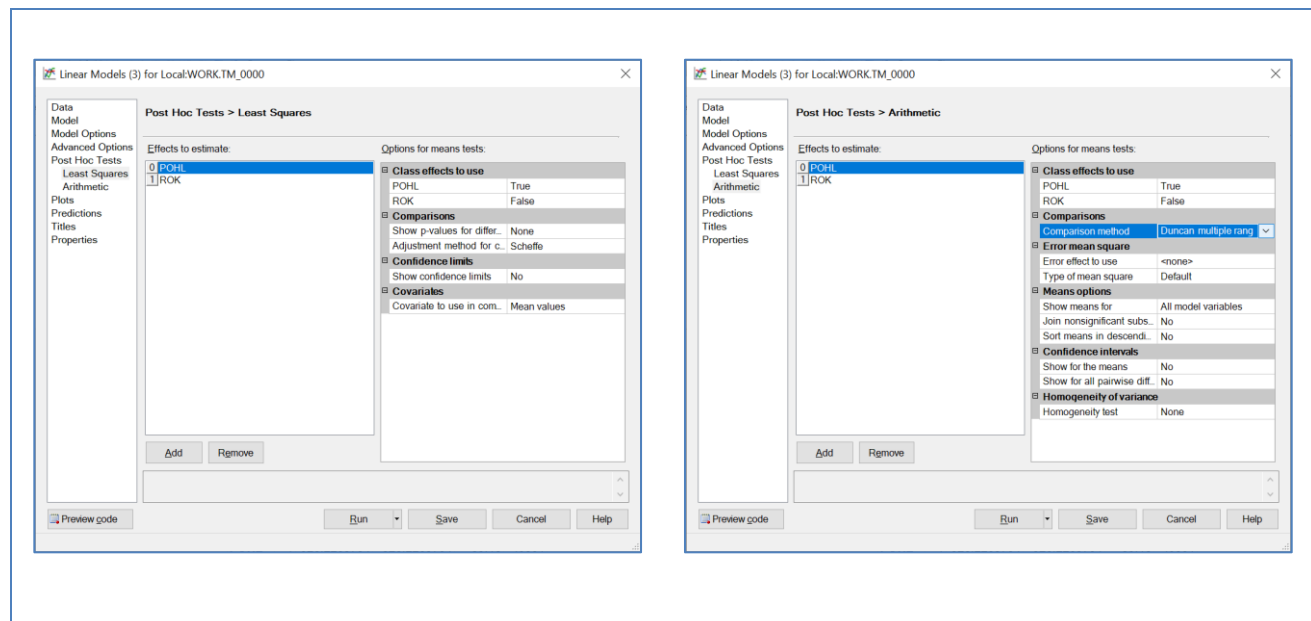
Na základe zobrazených grafov môžeme konštatovať, že ukazovateľ BMI má normálne rozdelenie aj v triedení podľa pohlavia študentov. Malá odchýlka od normality je zistená v prípade jej zobrazenia pri študentkách. Nepredstavuje ale výraznú odchýlku a preto podmienka zachovania normality údajov, potrebná pre analýzu rozptylu, je aj z tohto hľadiska splnená.

Nastavenie viacfaktorovej analýzy rozptylu:

Tab. 6.14a Nastavenie viacfaktorovej analýzy rozptylu



Tab. 6.14b Nastavenie viacfaktorovej analýzy rozptylu



Pri nastavení grafov bolo použité nastavenie viacerých vhodných grafov, ktoré ponúka program SAS Enterprise Guide.

Výsledky viacfaktorovej analýzy rozptylu a ich interpretácia

V tabuľke 6.14 uvádzame číselné výsledky viacfaktorovej analýzy rozptylu.

Tab. 6.14 Výsledky viacfaktorovej analýzy rozptylu

Class Level Information					
Class	Levels	Values			
POHL	2	M Z			
ROK	5	2008	2009	2011	2012 2014

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1005.413290	201.082658	19.36	<.0001
Error	483	5017.381429	10.387953		
Corrected Total	488	6022.794719			

R-Square	Coeff Var	Root MSE	BMI Mean
0.166935	14.27410	3.223035	22.57960

Source	DF	Type I SS	Mean Square	F Value	Pr > F
POHL	1	926.2269704	926.2269704	89.16	<.0001
ROK	4	79.1863194	19.7965798	1.91	0.1083

Source	DF	Type III SS	Mean Square	F Value	Pr > F
POHL	1	952.7105193	952.7105193	91.71	<.0001
ROK	4	79.1863194	19.7965798	1.91	0.1083

Zo základných číselných výsledkov viacfaktorovej analýzy rozptylu sa potvrdilo, že dva analyzované faktory spolu (pohlavie a rok) vysoko štatisticky ovplyvňujú variabilitu hodnôt BMI. F-test v modeli dosiahol hodnotu 19,36 ($p < 0,0001$). Faktory spolu vysvetľujú približne 17 % variability hodnoteného znaku BMI (koeficient determinácie dosiahol hodnotu 0,1669).

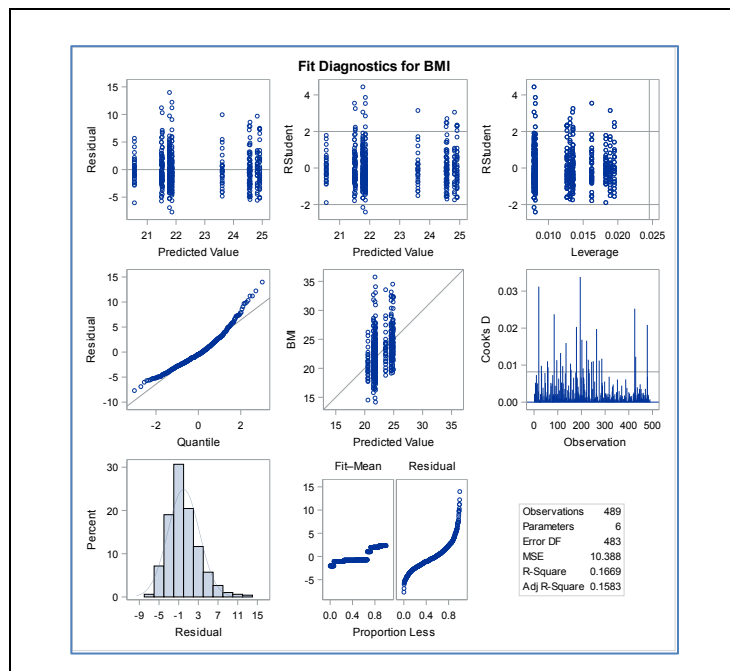
Nulová hypotéza o rovnosti priemerov podľa jednotlivých skupín faktorov sa nepotvrdila. Pri analýze jednotlivých faktorov sa potvrdila vysoká štatistická preukaznosť faktora pohlavie (POHL). Hodnota F-testu=89,16 resp. 91,71 ($p < 0,0001$).

Faktor rok (ROK) zostal štatisticky nepreukazný. Hodnota F-testu=1,91 ($p = 0,1083$). Analýza rozptylu umožňuje analyzovať jednotlivé faktory viacerými spôsobmi (4 typy výpočtu sumy štvorcov: Typ I, Typ II, Typ III a Typ IV).

V našej analýze sme použili Typ I a Typ III, ktoré sú štandardne nastavené v systéme (ich zmena je možná v rámci nastavenia výpočtu).

Grafické vyjadrenie vhodnosti (vhodná diagnostika) použitej viacfaktorovej analýzy rozptylu:

Tab. 6.15 Grafické výsledky viacfaktorovej analýzy rozptylu



Na základe jednotlivých grafov (hlavne grafy rezíduí) môžeme potvrdiť zachovanie normality rezíduí a opodstatnenosť nami použitej viacfaktorovej analýzy rozptylu.

Výsledky viacnásobných porovnávacích testov uvádzame iba pre faktor, ktorý v rámci analýzy dosiahol štatistickú preukaznosť (faktor pohlavie). Pri porovnaní LSM (Least Squares Means) priemerov sme použili Scheffeho viacnásobný test.

Tab. 6.16 Výsledky viacfaktorovej analýzy rozptylu - porovnanie LSM priemerov

		H0:LSMean1=LSMean2
POHL	BMI LSMEAN	Pr > t
M	24.4921195	<.0001
Z	21.4386605	

Scheffeho test potvrdil, že nulová hypotéza o rovnosti priemerov neplatí a preto platí alternatívna hypotéza, ktorá potvrdzuje, že LSM priemery podľa pohlavia sú vysoko štatisticky rozdielne ($p < 0,0001$).

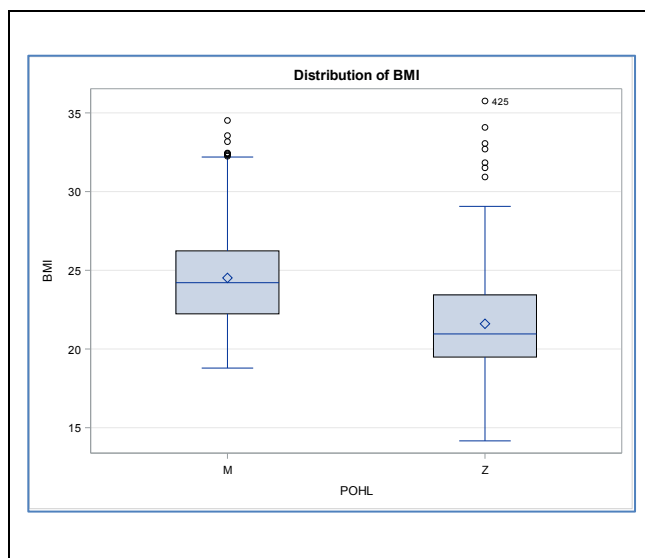
Pri porovnaní skutočných aritmetických priemerov sme tiež použili Scheffeho test.

Tab. 6.17 Výsledky viacfaktorovej analýzy rozptylu - porovnanie priemerov

Means with the same letter are not significantly different.			
Scheffe Grouping	Mean	N	POHL
A	24.5170	164	M
B	21.6019	325	Z

Na základe tohto testu sa tiež potvrdilo, že priemerné hodnoty BMI podľa pohlavia sú štatisticky signifikantne rozdielne (vytvorené dve rozdielne Scheffeho skupiny A a B). Uvedené je zrejmé aj z grafického znázornenia priemerov a variability znaku BMI podľa pohlavia študentov (tabuľka 6.18).

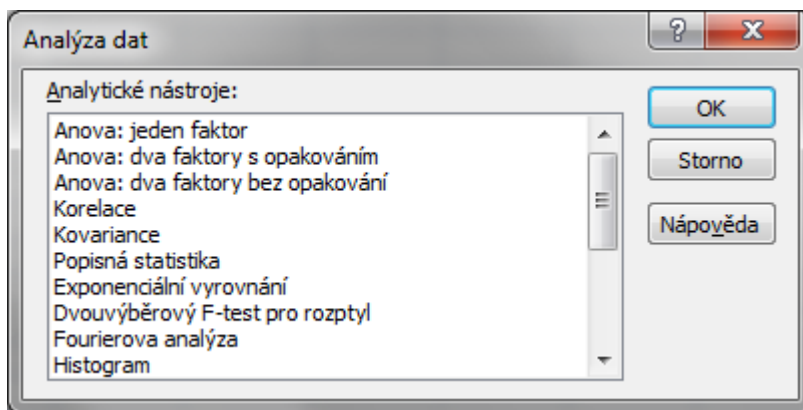
Tab. 6.18 Grafické výsledky viacfaktorovej analýzy rozptylu - porovnanie priemerov



Praktické použitie programu EXCEL (Microsoft 365)

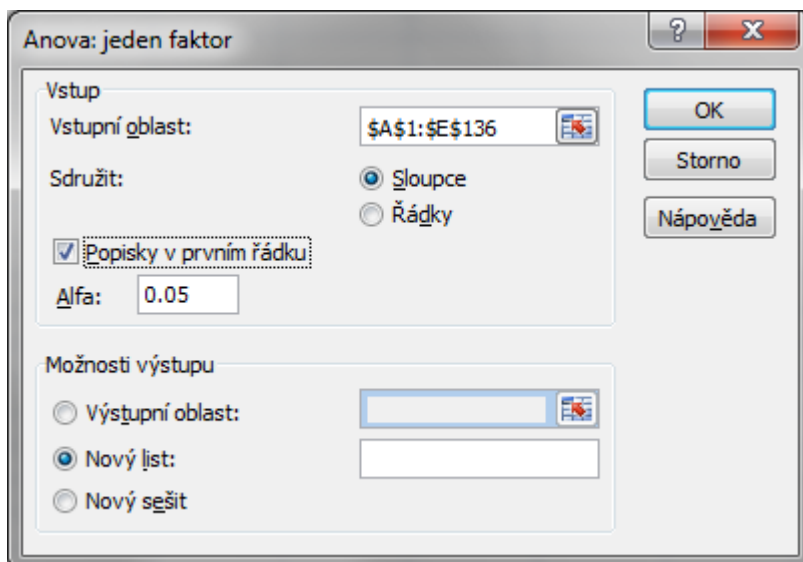
V rámci doplnkov programu Microsoft Excel je možné využiť doplnok analytických nástrojov **Analýza údajov** (Data Analysis), ktorý obsahuje niekoľko možností riešenia analýzy rozptylu: jednofaktorovú analýzu rozptylu a dvojfaktorovú analýzu rozptylu s opakovaním a bez opakovania. Zložitejšie viacfaktorové analýzy by sa ale riešili pomerne zložite a časovo zdĺhavejšie. V tabuľke 6.17 uvádzame možnosti doplnku analýzy údajov.

Tab. 6.17 Použitie analýzy údajov v programe Excel

**Príklad 1.1** (Excel) - jednofaktorová ANOVA

Detailné nastavenie vstupných (analyzovaných) údajov a odporúčané nastavenie ostatných parametrov výpočtu jednofaktorovej analýzy rozptylu uvádzame v tabuľka 6.18 (analyzovaný ukazovateľ musí byť umiestnený podľa úrovni hodnoteneého faktora vedľa seba v stĺpcoch).

Tab. 6.18 Nastavenie jednofaktorovej analýzy rozptylu v programe Excel



Výsledky (Excel)

V tabuľke 6.19 uvádzame výsledky jednofaktorovej analýzy rozptylu získané pomocou modulu Analýza údajov v programe Excel pre ukazovateľ BMI..

Tab. 6.19 Výsledky jednofaktorovej analýzy rozptylu (BMI)

Anova: Single Factor						
SUMMARY						
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
2008	62	1362.8	21.98064	9.819963		
2009	92	2126.369	23.11271	11.69002		
2011	66	1485.698	22.51057	12.92702		
2012	135	3060.536	22.67063	13.94987		
2014	134	3006.02	22.43299	12.01311		
ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	52.70277	4	13.17569	1.068164	0.371622	2.390359
Within Groups	5970.092	484	12.3349			
Total	6022.795	488				

Získané výsledky potvrdzujú nulovú hypotézu, že neexistuje štatisticky preukazný rozdiel medzi priemernými hodnotami ukazovateľa BMI (index telesnej hmotnosti) v skupinách študentov podľa rokov zisťovania telesných mier (hodnota F-testu 1,0681 neprekročila kritickú hodnotu F-testu 2,3903).

Rovnakú jednofaktorovú analýzu rozptylu sme uskutočnili aj pre ukazovateľ vek študentov. Pri tomto ukazovateli sa nepotvrdila nulová hypotéza o rovnosti priemerných hodnôt podľa rokov zisťovania telesných mier (hodnota F-testu 78,041 výrazne prekročila kritickú hodnotu F-testu 2,3903). Platí preto jednotková hypotéza o rozdielnosti priemerných hodnôt na úrovni 95 %.

Tab. 6.20 Výsledky jednofaktorovej analýzy rozptylu (vek)

Anova: jeden faktor						
Faktor						
<i>Výběr</i>	<i>Počet</i>	<i>Součet</i>	<i>Průměr</i>	<i>Rozptyl</i>		
2008	62	1267	20.43548	0.64331		
2009	92	1883	20.46739	0.932991		
2011	66	1498	22.69697	1.229837		
2012	135	3058	22.65185	1.094306		
2014	134	3052	22.77612	3.633711		
ANOVA						
<i>Zdroj variability</i>	<i>SS</i>	<i>Rozdíl</i>	<i>MS</i>	<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
Mezi výběry	537.9959	4	134.499	78.05417	4.62E-51	2.390359
Všechny výběry	834.0041	484	1.723149			
Celkem	1372	488				

Praktické použitie programu R (R Studio)

Príklad 1.1 (Program R) - jednofaktorová ANOVA

Príklad importu súboru TM.xlsx a zobrazenie údajov v rámci prostredia programu R Studio:

Import Excel Data

File/URL:
C:/Kega 2019-2021/Modul1/TM.xlsx Browse...

Data Preview:

PORADIE (double)	POHL (character)	VEK (double)	ROK (double)	HMOT (double)	VYSKA (double)	OBVODH (double)	OBVODP (double)	OBVODZ (double)	KRUZOK (double)	BMI (double)
1	Z	21	2008	52	160	90	70	15	1	20.1
2	Z	20	2008	54	174	86	68	16	1	17.4
3	Z	22	2008	57	174	83	63	16	1	18.1
4	Z	21	2008	62	168	100	76	15	1	21.1
5	M	21	2008	67	169	94	86	18	1	23.1
6	M	21	2008	72	192	94	89	17	1	19.1

Previewing first 50 entries.

Import Options:

Name: Max Rows: ☒ First Row as Names
 Sheet: Skip: ☒ Open Data Viewer
 Range: NA:

Code Preview:

```
library(readxl)
TM <- read_excel("C:/Kega 2019-2021/Modul1/TM.xlsx",
  sheet = "Data_TM")
View(TM)
```

[? Reading Excel files using readxl](#) Import Cancel

Zdrojový kód potrebný na import a zobrazenie excelovského súboru "TM.xlsx", ktorý je možné použiť priamo v programe R, alebo R Studio ako súbor s názvom **TM**:

```
# Import a zobrazenie údajov (databáza projektu KEGA)
```

```
library(readxl)
```

```
TM <- read_excel("C:/Kega 2019-2021/Modul1/TM.xlsx", sheet = "Data_TM")
```

```
View(TM)
```

```
# Zadanie popisnej štatistiky príkladu
```

```
# Inštalácia balíka summarytools (umožňuje veľkú rozmanitosť nastavení výpočtu)
```

```
install.packages("summarytools")
```

```
library(summarytools)
```

```
# Výpočet základných štatistických charakteristík podľa triediaceho znaku
```

```
stby(data = TM,
```

```
  INDICES = TM$ROK, # podľa roku
```

```
  FUN = descr, # descriptive statistics
```

```
  stats = "common") # most common descr. stats
```

Zadanie analýzy (program R)

```
# Analýza rozptylu (ANOVA) - jednofaktorová analýza
#
library(dplyr)
group_by(TM, ROK) %>%
  summarise(count = n(),
    mean = mean(BMI, na.rm = TRUE),
    sd = sd(BMI, na.rm = TRUE)
  )
#
boxplot(TM$BMI ~ TM$ROK)

boxplot(BMI~ROK,data=TM, main="BMI",
  xlab="Rok", ylab="BMI")

TM$ROK <- ordered(TM$ROK,
  levels = c(2008, 2009, 2011, 2012, 2014))

# Compute the analysis of variance
res.aov <- aov(BMI ~ ROK, data = TM)
# Summary of the analysis
summary(res.aov)

TukeyHSD(res.aov)

library(multcomp)
summary(glht(res.aov, linfct = mcp(ROK = "Tukey"))))

# 1. Homogeneity of variances
plot(res.aov, 1)

library(car)
leveneTest(BMI ~ ROK, data = TM)

# 2. Normality
plot(res.aov, 2)

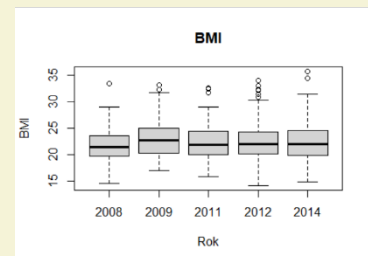
# Extract the residuals
aov_residuals <- residuals(object = res.aov )
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )
```

Výsledky analýzy rozptylu (program R)

Analýza rozptylu (ANOVA) - jednofaktorová analýza

#

	ROK	count	mean	sd
1	2008	62	22.0	3.13
2	2009	92	23.1	3.42
3	2011	66	22.5	3.60
4	2012	135	22.7	3.73
5	2014	134	22.4	3.47



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ROK	4	53	13.18	1.068	0.372
Residuals	484	5970	12.34		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = BMI ~ ROK, data = TM)

\$ROK

	diff	lwr	upr	p adj
2009-2008	1.13206433	-0.4480411	2.7121698	0.2864874
2011-2008	0.52992551	-1.1708721	2.2307231	0.9136223
2012-2008	0.68999090	-0.7853278	2.1653096	0.7033313
2014-2008	0.45234588	-1.0247043	1.9293961	0.9184880
2011-2009	-0.60213882	-2.1533757	0.9490981	0.8254604
2012-2009	-0.44207343	-1.7421460	0.8579991	0.8847704
2014-2009	-0.67971845	-1.9817556	0.6223187	0.6089413
2012-2011	0.16006539	-1.2842919	1.6044227	0.9981528
2014-2011	-0.07757963	-1.5237055	1.3685462	0.9998952
2014-2012	-0.23764502	-1.4103056	0.9350155	0.9813099

diff: difference between means of the two groups

lwr, upr: the lower and the upper end point of the confidence interval at 95% (default)

p adj: p-value after adjustment for the multiple comparisons.

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

```
Fit: aov(formula = BMI ~ ROK, data = TM)
```

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)
2009 - 2008 == 0	1.13206	0.57708	1.962	0.283
2011 - 2008 == 0	0.52993	0.62116	0.853	0.912
2012 - 2008 == 0	0.68999	0.53881	1.281	0.700
2014 - 2008 == 0	0.45235	0.53945	0.839	0.917
2011 - 2009 == 0	-0.60214	0.56654	-1.063	0.823
2012 - 2009 == 0	-0.44207	0.47481	-0.931	0.883
2014 - 2009 == 0	-0.67972	0.47553	-1.429	0.605
2012 - 2011 == 0	0.16007	0.52751	0.303	0.998
2014 - 2011 == 0	-0.07758	0.52815	-0.147	1.000
2014 - 2012 == 0	-0.23765	0.42828	-0.555	0.981

(Adjusted p values reported -- single-step method)

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	4	0.5254	0.7171
	484		

Shapiro-Wilk normality test

```
data: aov_residuals
W = 0.96007, p-value = 3.005e-10
```

Testovanie homogenity

